SPATIAL ANALYSIS OF UNCERTAIN THERMOBAROMETRIC DATA: APPLICATION TO THE SWISS CENTRAL ALPS

Tivadar M Tóth[1,2]

[1] Department of Mineralogy, Geochemistry and Petrology, University of Szeged, Hungary (mtoth@geo.u-szeged.hu)

Abstract

Several approaches exist to estimate peak pressure and temperature (PT) conditions of a single metamorphic rock sample. Because of many reasons from analytical problems to uncertainty in mineral solution models, all these calculations are rather uncertain making spatial interpretation of the data set for a whole metamorphic terrain problematic. In this study Error Kriging (de Marsily, 1984) is used to calculate PT maps for the Swiss Central Alps and the results are compared to those got by other kriging methods.

Keywords: thermobarometry, uncertainty, error kriging, Central Alps.

1. INTRODUCTION

When one seeks to interpret the results of thermobarometric studies, the spatial context is obviously important. Graphical representations, e.g. isotherms and isobars on maps or profiles, are helpful in interpreting thermobarometric data sets in a regional context. For geologists to assess the significance and implications of P-T data, it may indeed be crucial to see them together with results from geophysical, tectonic, or geochronological studies. Similarly, in modeling studies it may be most useful to compare the thermal or baric field of an orogen with predictions based on theoretical simulations. For all these reasons, we need reliable graphical representations of spatially discrete P-T data sets. As these data sets may include various forms of information, commonly containing quite variable uncertainties, the task of interpolating and extrapolating such data sets demands adequate tools.

2. PT data

Several methods exist to estimate P and T data for a given metamorphic assemblage, either from traditional Fe-Mg exchange thermometers and net-transfer type barometers (such as GASP), or from more reliable multi-equilibria techniques (Berman 1991, Gordon 1992). A

number of authors (e.g. Kohn, Spear 1991, Lieberman, Petrakakis 1991, Gordon 1992, Holland, Powell 1994) have investigated the uncertainty of various PT-estimation methods, which are due to numerous sources of error, like electron microprobe calibration, counting statistics during measurement, mineral formula calculation, mineral solution models among many others. Evaluation of these individual uncertainties may allow their propagation into a finite statistical error of P and T data. Such results may then be quoted as „crisp" PT data (e.g. 500±25°C, 2.8±0.6 kbar). Other thermobarometric approaches rely on phase diagrams (such as petrogenetic grids calculated using DOMINO/THERIAK software) and produce a "permitted" PT-window for each assemblage (e.g. 500-540°C). In this case no statistical error value is calculated; uncertainty appears as an inequality type of datum, i.e. an interval in T or P.
A different way to present uncertainty is by means of fuzzy numbers (e.g. Dubois, Prade, 2000). Fuzzy numbers are defined through "membership functions". The value of a membership function (m(x) in the [0,1] interval) depends on how *possible* (not how mathematically probable!) the datum (x) is. Defining the fuzzy numbers is usually based on qualitative or semiquantitative information. For example, if temperature is with known (without any doubt) to lie between 450 and 600 °C, and is possibly in the 500-550 °C interval, the proper fuzzy number would be a trapezoid:

$$m(T) = \begin{cases} 0, & \text{if } T < 450\ ^\circ C \\ (T-450)/50, & \text{if } 450\ ^\circ C < T < 500\ ^\circ C \\ 1, & \text{if } 500\ ^\circ C < T < 550\ ^\circ C \\ (600-T)/50, & \text{if } 550\ ^\circ C < T < 600\ ^\circ C \\ 0, & \text{if } T > 600\ ^\circ C \end{cases}$$

There are two reasons to prefer fuzzy numbers to standard error. Firstly, there is in most cases no proof that the uncertainty in P- and T-data are of probability type (e.g. Gaussian). Secondly, both error and inequality type data are easily transformed to fuzzy data, which thus offer a way to construct a data set having a uniform measure of uncertainty.
Although, there are different ways to define and calculate uncertainty in data, their spatial representation is more problematic; interpolating between data that have different uncertainty is not a simple matter. In what follows, different interpolation methods are presented and tested, which are considered promising for mapping P- and T-data.

3. Interpolation methods applied

Thermobarometry yields P and T data in spatially discrete form, hence for many questions spatial interpolation (or limited extrapolation) is required, most notably in the production of maps and profiles. Traditional interpolation techniques, such as linear interpolation (by hand or machine), trend surface analysis or inverse distance interpolation have long been found useful, yet they all have significant disadvantages. Inverse distance method, for example, tends to generate unrealistic „bull's-eye" shaped structures surrounding the position of data points. None of them take into account the real spatial structure of the data set, and none of them allow an estimation of interpolation error.

Kriging, a family of stochastic interpolation methods, has a fundamental role in geostatistics for decades. Various texts and handbooks go deep into its basic concept and are not discussed here (e.g. Matheron, 1970, Cressie, 1991), Wackernagel, 1995). In what follows ordinary kriging system (OK) with and without nugget effect will be used.

4. Kriging with uncertain data

Ordinary kriging is a good interpolator in cases where reliable data exist in sufficient number (and spatial density). However, geostatistics has spread into several areas including hydrology and soil sciences, where the conditions and requirements for OK are not always satisfied. A common problem is the insufficient quantity or quality of measurements. In order to get a useful interpolator in such cases, a number of attempts have been made to incorporate additional information into the kriging system. Bárdossy et al. (1988) present an exhaustive collection of these approaches. Some of these shall be discussed here in the context of thermobarometry, the goal being to represent regional results as maps of continuous isotherms and isobars.

In the case of mapping metamorphic P and T, difficulties arise because the data are of variable type and precision, and their spatial distribution tends to be far from uniform. Methods used to estimate statistical errors in thermobarometric data have received attention in recent years (e.g. Kohn, Spear 1991, Lieberman, Petrakakis 1991, Gordon 1992, Holland, Powell 1994). Where the data set suggests that other types of uncertainty should be incorporated in the kriging system, this is possible by two approaches called *soft kriging* (Journel, 1986) and *fuzzy kriging*. A soft kriging system works with inequality type data and/or constraint intervals. This procedure may be useful in PT mapping, where the mineral assemblages or petrogenetic net model allow only an estimation of P-T intervals (e.g. T_{min} and T_{max}). Both error and interval type uncertainties can easily be transformed to fuzzy numbers, making it possible to use data sets with mixed types of information. However, at the present stage, the fuzzy kriging estimator tends to use either only fuzzy or only crisp data for interpolating, depending on the initial conditions (Bárdossy et al., 1990a, b). For fuzzy kriging to yield reliable results, considerable computational effort is needed.

Error kriging [1] (EK) follows very simply from OK: One uses an error ε_i associated with each datum $Z(x_i)$, with the following constraints (Marsily, 1984):

- $E[\varepsilon_i] = 0, i = 1 \ldots n$ ε_i is not a systematic error. (E is the mean);
- $Cov[\varepsilon_i, \varepsilon_j] = 0, \forall\, i \neq j$ errors (ε) are not correlated with each other;
- $Cov[\varepsilon_i, Z(x_j)] = 0, \forall i, \forall j$ errors (ε) are not correlated with data;
- σ_i^2 is known for each i.

Compared to OK, the only difference in these conditions is that the kriging system now has values of $-\sigma_i^2$ (instead of zero) in the diagonal elements of the error matrix. As a consequence, the estimator uses normal distributions with a given mean and variance for interpolation. It is probably fair to say that only the best thermobarometric data sets may come close to satisfying the above four conditions reasonably well. For example, even if multi-equilibrium techniques are used, a (minor) systematic error cannot be ruled out. Additionally, at low temperature

[1] Following a suggestion by A. Bárdossy (pers. comm.), the term "kriging with uncertain data" used by Marsily (1984) has been changed to "error kriging".

conditions T estimation becomes less accurate, hence there is likely a slight correlation between T-data and their error. However, for good thermobarometric data sets, estimation errors should come close to satisfying all four constraints fairly closely, in which case EK is the method of choice to interpolate P-T data.
The use of OK and EK kriging to generate PT maps is presented here for the example of the Swiss Central Alps. By way of example, the main steps required in any type of kriging analysis and the graphical representation of the results are documented first. In the subsequent section the results are tested by cross validation (e.g. Wackernagel, 1995), and the maps obtained by these two different approaches are compared.

5. A case study: P, T maps of the Swiss Central Alps

120 T-data and 97 P-data for the Swiss Central Alps have been selected for constructing isotherms and isobars. For a description of the geology, methods, and data sets see Engi et al. (1995) and Todd, Engi (1997). Error data in P and T were calculated using program INERSX, if at least three independent reactions existed. For all other data points T_{err} was assumed as 50°, and P_{err} = 2 kbar.

5.1. Sequence of spatial data analysis

The spatial distribution of data points in both cases (P and T) is rather uneven. In relatively large areas, geological conditions are unfavorable, and no data points exist, whereas clusters of data exist in valleys with suitable outcrops, where detailed and multiple sampling was possible. Clustered distributions of data points usually lead to an estimation error in variography (Armstrong, 1984). In order to avoid this problem, a moving window declustering method was used prior to the calculation of variograms. This process substitutes data that fall into a given rectangle (spatial window) by their arithmetic mean. Moving this window and calculating the means over the study area result in a data distribution of equal density. The size of the windows is chosen on the basis of the average spacing between locations and the size of the entire area being studied. If the window is too large, the number of points left after the process insufficient. If it is chosen too small, no reliable statistics can be obtained for most windows. For the area studied in the Central Alps (~4000 km^2), overlapping windows 5*5 km in size were found satisfactory for both T and P. For moving window statistic, calculations the program MWINDOW (Murray, Baker, 1991) was used. Variograms were calculated based on the cluster means rather then the original database. Both experimental variogram calculations and theoretical variogram fitting were performed using the program VARIOWIN (Pannatier, 1994).

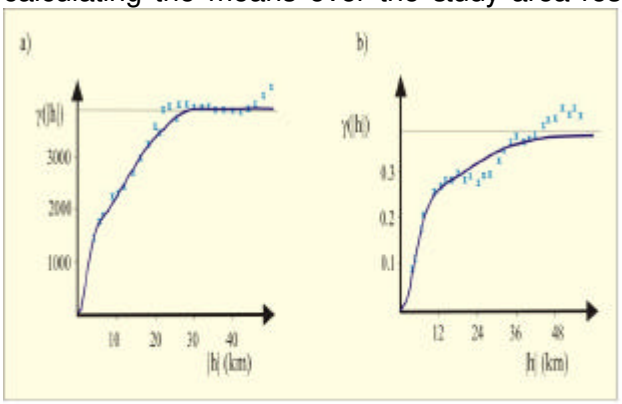


Fig.1 Estimated variograms for a) temperature, b) pressure for the Swiss Central Alps. For best-fit models see text.

On the experimental variogram that characterizes the spatial variation in temperature data, two different sills can be distinguished. The nested structure variogram (Serra, 1968) fitted to it is the sum of two Gaussian type variograms. One of them has small (5 km) range, the other

significantly larger (29.5 km) range: $\gamma(h)=1520*G(5)+2520*G(29.5)$, where "G" denotes the Gaussian variogram function (***Fig. 1/a***). A sum of two Gaussian variograms was also found best for pressure (***Fig.1/b.***) ($\gamma(h)=0.234*G(10.2)+0.148*G(43.6)$), but the range of both individual variograms for P is significantly higher than in the case of T. The fit for P is weaker than for T, probably due to a hole effect (Cressie, 1991), shown by a negative correlation in a small range of the variogram (***Fig. 1/b***). Anisotropy was not calculated for either case.

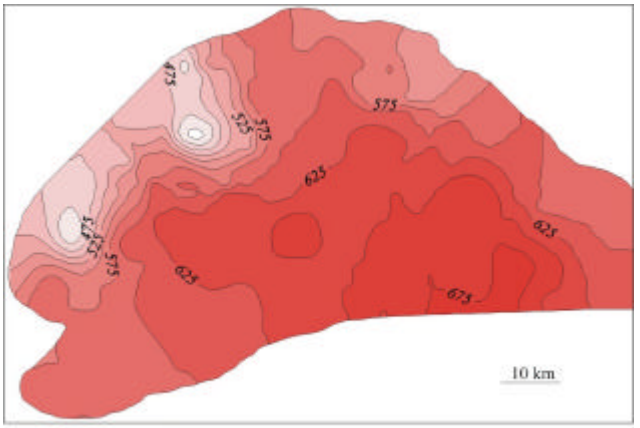


Fig.2/a Calculated PT maps for the Swiss Central Alps using various interpolation methods. EK map for temperature (T in °C)

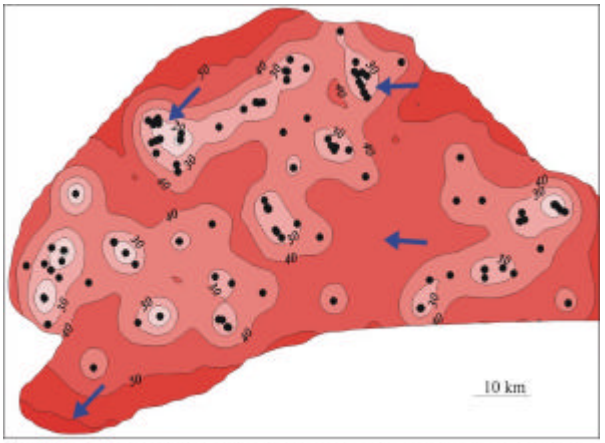


Fig.2/b Calculated PT maps for the Swiss Central Alps using various interpolation methods Kriging variance for the EK map of temperature (T in °C) for arrows see text)

Finally, theoretical variogram parameters were combined with the original data for OK and EK in a program written by Bárdossy, A. (unpublished). This code allows both data estimation and kriging error calculation to be performed simultaneously. Each kriging parameter was chosen to be the same in the two processes, but for OK the estimated errors were set zero. Contour lines were then generated using program SURFER (Golden Software Inc., 1994) (***Fig. 2/a-g***).

5.2. Results and Comparison of Different Methods

EK error maps for both P and T exhibit low kriging standard deviations (***Fig. 2/a-d***). In both cases the interpolation is best where data points form clusters, while the error is largest in areas for which no data exist (see arrows on ***Figs 2/b, 2/d***). Although the structure of the two error maps is similar, the dimension of low error areas is significantly smaller for pressure (***Fig. 2/d***). In addition, data clusters with a large error exist, indicating that the uncertainties in metamorphic pressure estimates are larger than in T.

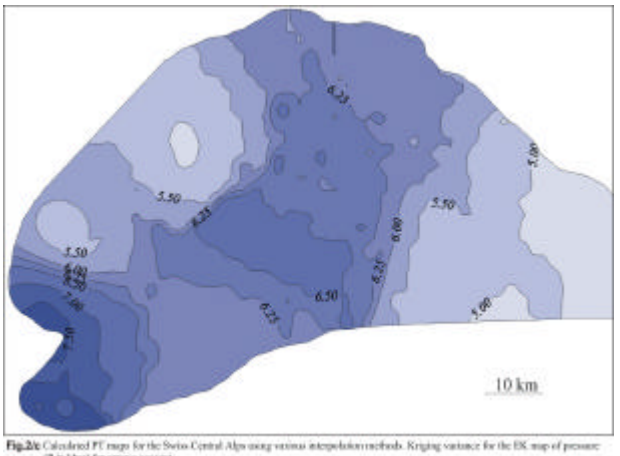


Fig.2/c Calculated PT maps for the Swiss Central Alps using various interpolation methods. Kriging variance for the EK map of pressure (P in kbar) for arrows see text

OK maps for both T and P exhibit unrealistic results, with peaks and ditches forming in many areas. Extreme values in P – up to 20 kbar or as low as 0 kbar – were locally interpolated (***Fig. 2/e***). Estimated temperature data vary

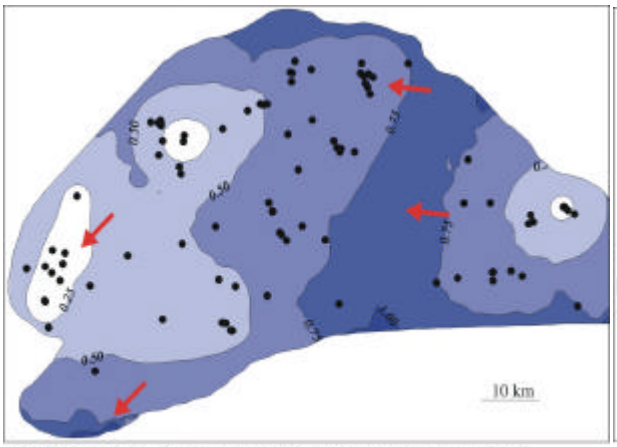


Fig.2/d Calculated PT maps for the Swiss Central Alps using various interpolation methods. EK map for pressure (P in kbar)

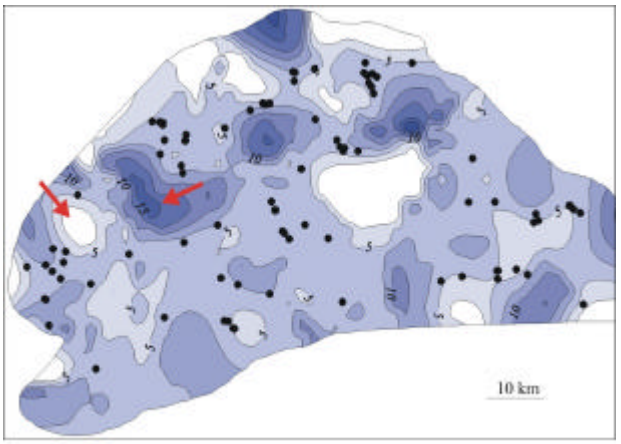


Fig.2/e Calculated PT maps for the Swiss Central Alps using various interpolation methods. OK map for temperature (T in °C)

between 0 and 1700 °C (not presented). The results emphasize the problem of using exact interpolation methods for data sets that contain uncertainty. The reason of the conspicuous (mis)estimation is likely that significantly different values, both in the initial P and T data sets, exist close to each other. Because no nugget effect was involved in the variogram used, OK is an exact interpolator, i.e. the estimated surface tends to pass through each datum point. This ambition leads to bad interpolation results in areas where significantly different values exist. If OK with a nugget effect model (smoothing interpolator) is used, a more realistic result may be obtained. By using this variogram for kriging, one can generate smoothed maps even in areas having higher uncertainty in the data (***Fig. 2/f***). The error of estimation using OK is significantly larger than for EK in areas where data points are numerous (***Fig. 2/g***). On the other hand, for extrapolation OK appears to be more reliable, as shown by the small error values towards the border of the study area as well as in domains lacking data.

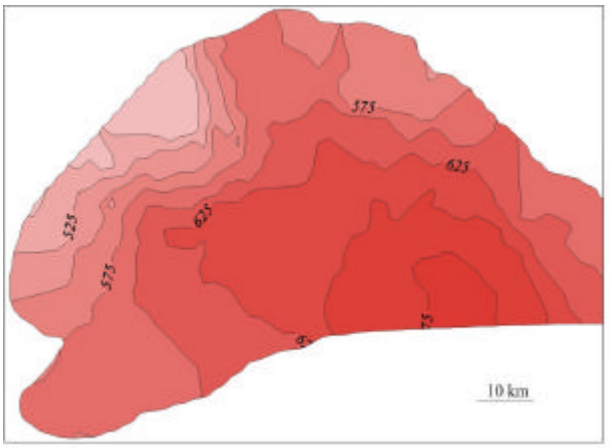


Fig.2/f Calculated PT maps for the Swiss Central Alps using various interpolation methods. OK map (with nugget effect) for temperature (T in °C)

The reliability of two maps calculated by EK and OK with nugget effect, respectively, can also be tested by cross a validation procedure. In this process each sample value at location x_0 is removed in turn from the data set and $Z^*(x_0)$ is estimated using the other samples. Comparison of the measured and the estimated data helps to identify apparent bias. For EK and OK the average of the T differences is as low as 2.5 and 0.8 °C, respectively, no systematic overestimation is seen in either case. Correlation coefficients

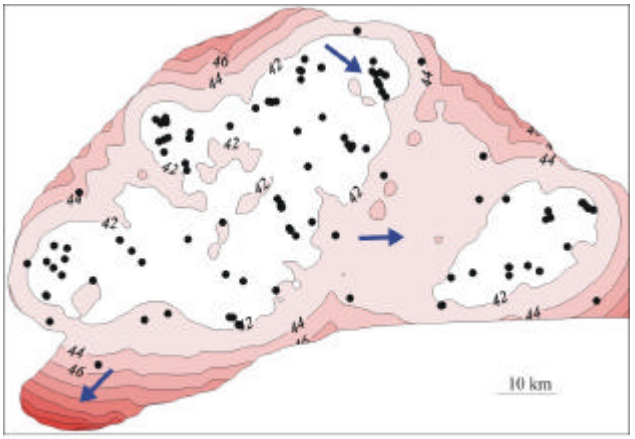


Fig.2/g Calculated PT maps for the Swiss Central Alps using various interpolation methods. Kriging variance for the OK map (with nugget effect) of temperature (T in °C); for arrows see text.

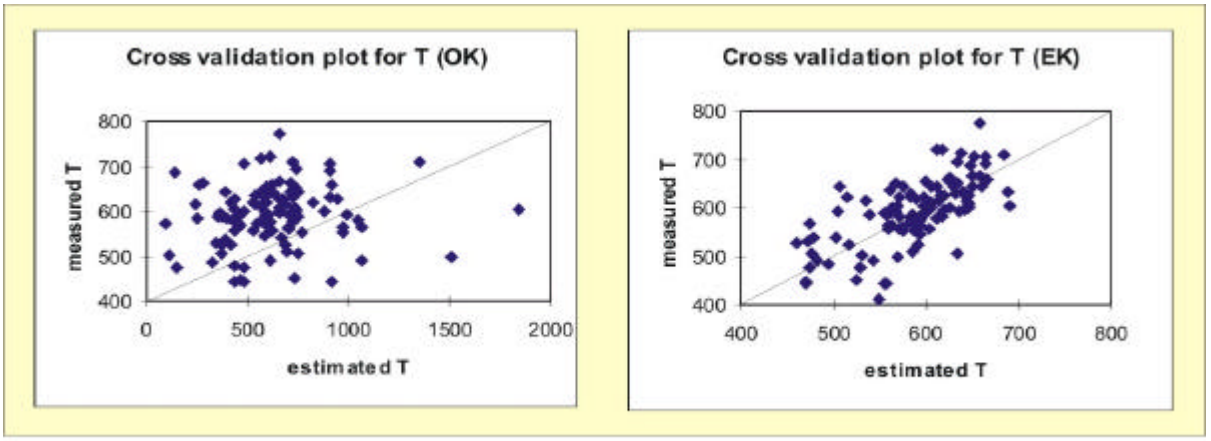


Fig.3 Cross validation plots for T (°C) calculated by a) EK and b) OK respectively. For details see text.

between estimated and measured data are similar in the two cases (r=0.67 for both), and the two estimations produce similar data (r=0.99). On ***Fig. 4*** one can see that EK reproduce the original data well, while the measured and the estimated T data are barely correlated (r=0.14) in the case of the OK map (without nugget effect). Whether the slight underestimation for high T data and the tendency to overestimate low T data (***Fig. 3***) for both EK and OK is characteristic of these interpolation methods, or whether these effects are inherited from the data set used is out of the scope of the present study.

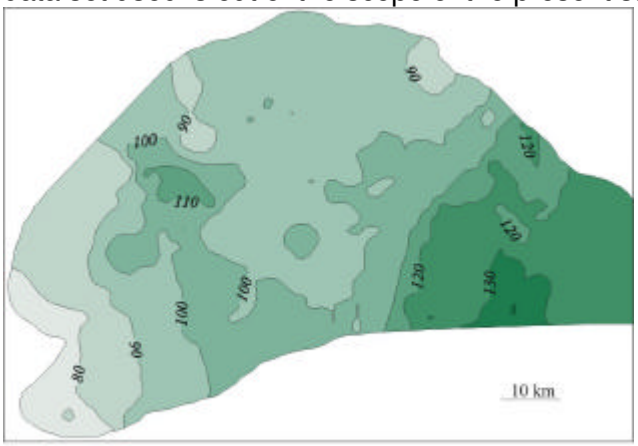


Fig.4 Metamorphic T/P (°C/kbar) map for the Swiss Central Alps

Based on the previous reasoning one can say that reliable maps can only be obtained if the uncertainty in the data is involved in the interpolation algorithm. The two approaches used (EK, OK) have different advantages and disadvantages. If a sufficient number of data points exist, with a spatial distribution close to uniform, EK is preferred because this method produces the smallest kriging error. For highly clustered data sets, as well as to get reliable maps close to the border of the area, OK is the superior method.

5.3. Geological representation of kriging results

For possible application of the PT maps in geological interpretation two examples are briefly presented. In both cases maps got by EK method are used.

1) The T/P ratio may be an informative parameter when comparing different metamorphic terrains. Based on isotherms and isobars, the construction of a T/P map is a simple calculation (***Fig. 4***). To propagate errors into the T/P map, the following expression should be used:

$$\sigma_{T/P} = \sqrt{\sum_{P}\sum_{T}\left(\frac{\partial(T/P)}{\partial P}\right)*\left(\frac{\partial(T/P)}{\partial T}\right)*\sigma_P*\sigma_T*r_{PT}}$$

where r_{PT} is the correlation coefficient between P and T.

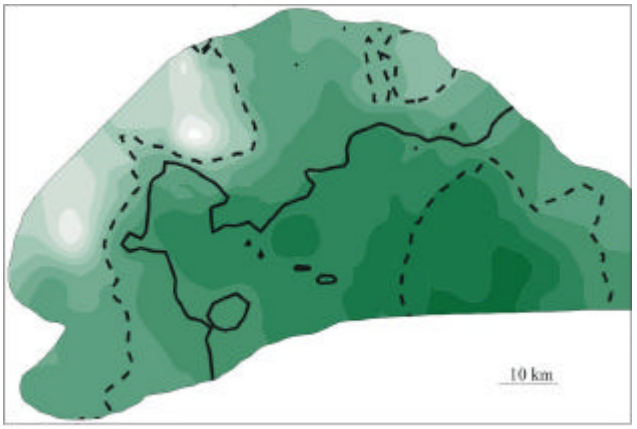


Fig.5 Calculated sillimanite-in isograd for the Swiss Central Alps. Dashed line shows the 90% probability envelope

2) The exact position of the sillimanite$_{in}$ isograd in the Central Alps has been argued for a long time. By taking into account the experimentally determined P-T-location (e.g. Holdaway, 1971) of the kyanite-sillimanite reaction, the appropriate combination of isotherms and isobars results in the representation of the univariant curve in the real space. If the error maps are considered as well (by using a function equivalent to the one above), one can represent the stability limit for sillimanite with an error envelope (***Fig. 5***). This confidence interval may then be compared to the mineral zone boundary as delimited by field observation (e.g. Irouschek 1982, Todd, Engi 1997).

Acknowledgments

Discussions with A. Bárdossy on methods in spatial statistics, and the availability of his computer code for kriging have helped this project. His interest and contributions are gratefully acknowledged. M. Engi and C. S. Todd are thanked for their valuable remarks and for cleaning the English. The Bolyai Research Grant supported the project.

References

Armstrong, M. (1984): Common problems seen in variography. Mathematical Geology, **16/3**, 305-313.

Bárdossy, A., Bogárdi, I., Kelly, W. E. (1988): Imprecise (fuzzy) information in geostatistics. Mathematical Geology, **20**, 287- 311.

Bárdossy, A., Bogárdi, I., Kelly, W. E. (1990): Kriging with imprecise (fuzzy) variograms. I. Theory. Mathematical Geology, **22**, 63-79.

Bárdossy, A., Bogárdi, I., Kelly, W. E. (1990): Kriging with imprecise (fuzzy) variograms. II. Application. Mathematical Geology, **22**, 81-95.

Berman, R. G. (1991): Thermobarometry using multi-equilibrium calculations: a new technique, with petrological applications. Canadian Mineralogist, **29**, 833-855.

Cressie, N. A. (1991): Statistics for spatial data. Wiley & Sons, New York, 900 p.

Dubois, D., Prade, H. (2000): Fundamentals of fuzzy sets. Kluwer Academic Publishers

Engi, M., Todd, C. S., Schmatz, D. R. (1995): Tertiary metamorphic conditions in the eastern Lepontine Alps. Schweizerische Mineralogisches Petrographisches Mitteulungen, **75**, 347-369.

Golden Software Inc. (1994): SURFER for Windows User's guide. Golden Software Inc., Golden, Colorado, variously paged

Gordon T. M. (1992): Generalized thermobarometry: Solution of the inverse chemical equilibrium problem using data for individual species. Geochimica Cosmochimica Acta, **56**, 1793-1800.

Holland T. J. B., Powell R. (1994): Optimal geothermometry and geobarometry. American Mineralogist, **79**, 120-133.

Irouschek, A. (1983): Mineralogie und Petrographie von Metapeliten der Simano-Decke unter besonderer Berücksichtigung cordieritführender Gesteine zwischen Alpe Sponda und Biasca. Unpublished PhD Dissertation, University of Basel, Switzerland.

Isaaks, E. H., Srivastava, R. M. (1989): Applied geostatistics. Oxford Univ. Press, New York, 544 p.

Journel, A. G. (1986): Constrained interpolation and qualitative information - The soft kriging approach. Mathematical Geology, **18/3**, 269-286.

Kohn M. J., Spear F. S. (1991): Error propagation for barometers: 2. Application to rocks. American Mineralogist, **76**, 138-147.

JOURNAL OF HUNGARIAN GEOMATHEMATICS

Volume 1

Geomathematics in Hungarian Geology

George Bárdossy[1]

[1]Hungarian Academy of Sciences (h4750bar@helka.iif.hu)

Abstract

The application of mathematical methods has a long tradition in Hungary. The main bases of geomathematics are the universities of the country, more closely the departments related to geology, such as general geology, stratigraphy, paleontology, structural geology, mineralogy, pet-rography, geochemistry, hydrogeology and applied geology. The Hungarian Geological Survey, the Geological Institute of Hungary and the Geochemical Research Laboratory of the Hungarian Academy of Sciences are institutions where geomathematical methods found broad applications. Finally, some mining and exploration companies, like the Hungarian Oil Company (MOL), the Bakony Bauxite Mining Company and others are regularly using geomathematical methods, mainly for the evaluation of exploration results, for deposit and reservoir modelling and for the estimation of resources.

Keywords: geomathematics, applications.

The application of mathematical methods has a long tradition in Hungary. The main bases of geomathematics are the universities of the country, more closely the departments related to geology, such as general geology, stratigraphy, paleontology, structural geology, mineralogy, pet-rography, geochemistry, hydrogeology and applied geology. The Hungarian Geological Survey, the Geological Institute of Hungary and the Geochemical Research Laboratory of the Hungarian Academy of Sciences are institutions where geomathematical methods found broad applications. Finally, some mining and exploration companies, like the Hungarian Oil Company (MOL), the Bakony Bauxite Mining Company and others are regularly using geomathematical methods, mainly for the evaluation of exploration results, for deposit and reservoir modelling and for the estimation of resources. In the following examples of recent geomathematical applications are listed.

Eötvös Loránd University, Department of Applied Geology, Budapest

Under the leading of Dr. J. Kovács:

- ***Time- trend analyses*** of underground water systems and sedimentary sequences

- ***multivariate statistical methods*** e.g. cluster- ,principal component-,dynamic factor analyses for the study of underground hydrogeologic systems
- ***spatial analyses*** of underground water systems e.g. autoeregression, moving averages, point- and block kriging

Directed by Dr. A. Füst:

- ***geostatistical calculations*** on different types of mineral deposits in Hungary, deposit modelling.

Szent István University, Institute of Mathematics and Informatics, Gödöllõ

Directed by Prof. Z. Varga and Dr. Z. Sebestyén:

- Statistical evaluation of different sedimentary sequences by ***bivariate and multivariate statistical methods*** and by ***geostatistics***, mainly ***variography***
- Application of the ***bootstrap method*** for the statistical evaluation of small sized samles (N=10-30)

Szent István University, Department of Biomathematics and Informatics, Budapest

Directed by Prof. J. Fodor:

- application of ***fuzzy systems*** and ***fuzzy logic*** to the ***treatment of uncertainties and errors*** in geology,e.g. quantitative phase analysis of rocks by X-ray diffractometry and thermal methods, safety assessment of radioactive waste repositories, transmissivity measurements of underground water in boreholes.

University of Szeged, Department of Mineralogy, Geochemistry and Petrography, Szeged

Under the leading of Dr. T. M. Tóth:

- modelling of fractured crystalline rocks by the methods of ***fractal geometry***.

University of Szeged, Department of Geology and Paleontology, Szeged

Under the leading of Dr. J. Geiger:

- ***3D geostatistical modelling*** of ancient fluvial dominated delta environments.
- ***Time series analysis*** of sedimentary micro-cycles based on Computer Tomograph data of hand specimens
- ***Multivariate statistical methods*** for identification of ancient sedimentary environments from grain size data and sedimentary structures
- ***Markov chain analysis*** of sedimentary sequences.

Miskolc University, Miskolc

Prof. F. Steiner:

- Detailed ***theoretical investigation of robust estimators***. Practical applications in geology and geophysics.

Hungarian Geological Survey, Budapest

- evaluation of magmatic rocks in Hungary by ***various statistical methods*** **(L. Ó. Kovács and G. Kovács)**
- ***application of the fuzzy set theory*** to the resource estimation of solid mineral deposits in Hungary (**B. Fodor and G. Szebényi**), modernization of the traditional resource estimation methods.

Geological Institute of Hungary, Budapest

- investigation of the ***background concentrations of chemical elements*** of rocks and soils in Hungary ***by statistical methods*** (**I. Horváth, P. Scharek**)
- ***geochemical evaluation*** of chemical data of soils in Hungary ***by statistical methods*** for agro-geological purposes (**U. Fügedi**)
- statistical evaluation of hydrogeological data systems, including ***time-trend analyses*** (**Á. Szalkay, I. Horváth, Gy. Tóth**)
- ***Statistical evaluation of micro-tectonic data***, calculation of paleo-stress (**Gy. Maros, K. Palotás, L. Fodor, B. Koroknai**)
- Modelling of fracture systems in rocks by methods of ***fractal-geometry***, ***detection of fault systems by statistical methods*** (**Z. Unger**)
- Geological ***evaluation of aerial and satellite pictures*** (**Z. Unger**)

Geochemical Research Laboratory of the Hungarian Academy of Sciences.Budapest

Directed by P. Árkai:

- ***statistical evaluation*** of measures of crystallinity of layer silicate minerals, ***application of fuzzy set theory*** to the evaluation of the results of quantitative mineralogical phase analyses.

Hungarian Oil Company(MOL), Szeged

Under the leading of J. Geiger:

- ***Variography***
- Numerical modelling of the properties of oil and gas reservoirs by ***indicator kriging*** and ***co-kriging, sequential Gaussian simulation, Markov-Bayes simulation, turning band simulation***.
- ***2D and 3D spatial modelling*** of reserves.
- ***parameter upscaling*** for dinamic flow-models and ***study of the volume-effect***.
- Study of the ***spatial uncertainty*** of the models.

Bakony Bauxite Mining Company,Tapolca

- Resource estimation of bauxite deposits by traditional methods and by the ***application of the fuzzy set theory***.(**I. R. Szabó, G. Varga**).
- Geostatistical calculations for mining-geological purposes: ***variography, point- and block kriging*** (**S. Diószegi**).

Finally**, the author of this paper** is studying the ***uncertainties and errors of geological investigations and possibilities of their evaluation by traditional stochastic and some new mathematical methods, e.g fuzy set theory, fuzzy logic, probability bounds.(Hungarian Academy of Sciences, Department of Earth Sciences)***
Applications of mathematical methods in other branches of the Earth Sciences, e.g. geography, geophysics, geodesy, mining etc.will be outlined in separate papers.Those who would like to have more informations about some of the investigations listed above, are requested to send an e-mail message to George Bárdossy (h4750bar@helka.iif.hu).

A PRECAUTIONARY GROUNDWATER POLLUTION POTENTIAL ASSESSMENT AIDED BY A FUZZY LOGIC SYSTEM

E. Cameron[1], G. F. Peloso[2]

[1]GeoStudio, P.zza S. Antonio 15, 23017 Morbegno (SO), Italy.

Tel.: +39 0342 615987, E-mail: geostd@libero.it

[2]Universitàdegli studi di Pavia, Dipartimento di Scienze della Terra, Via Ferrata 1, 27100 Pavia, Italy.

Tel.: +39 0382 505828, E-mail: gpeloso@unipv.it

Abstract

This paper presents a fuzzy logic based approach to data uncertainty management in groundwater pollution potential assessment; a modified parametric model produces a pollution potential score that ensures a degree of groundwater protection which increases with parameters measurement or estimation uncertainty. The model behavior is investigated and the consequences of its use are outlined with respect to risk analysis.

Keywords: fuzzy logic, data uncertainty management, groundwater pollution potential assessment, SINTACS, f-SINTACS

1. PREMISES

Aquifers' pollution potential assessment is an important planning and pollution prevention tool.
Among the assessment methods DRASTIC (Aller et. al. 1987) and SINTACS (Civita and De Maio, Italian Research Council, 1997) assign a partial pollution potential score to each value of the parameters assumed to be relevant (groundwater table depth, net recharge etc.) and produce a total pollution potential score as a weighted sum of the partial ones, with weights that depend on the specific hydrogeological situation.
The actual parameters values are generally not known exactly, mainly because of measurement or estimation errors, and the same apply, therefore, to the true score. This raises the problem of choosing an appropriate pollution potential level for planning and design decisions.
The problem can simply be ignored using only the measure or estimate of each parameter, obtaining directly the final score; it can be argued, however, that because this last is affected by some degree of uncertainty the real pollution potential at a given site may be underestimated, and therefore planning and design decision might not be sufficiently safe for groundwater resources.
Parameter data quality can be increased but time, costs and different technical limitations may not allow to reach the desired level of accuracy.
The model presented in this paper attempts to introduce a safety degree in pollution potential assessment with a given data set; a fuzzy logic system calculates a partial score starting from a parameter measure or best estimate and the mathematically formulated notions of acceptable approximation to its actual value and of negative influence of the parameter itself on pollution potential level.
The model is expected to introduce a quantified and reasonable degree of precaution in pollution potential assessment, reducing the possibility to take planning and design decisions not sufficiently safe for groundwater resources. The underlying precaution principle receives the common interpretation which suggests to take decisions with a safety level directly related to the uncertainty of the information on which they are based; there is no radical or prejudicial action restraining approach and the procedure, which will now be briefly outlined, is mathematically defined even in its subjective elements.
Let's assume that a_P is the actual value of a given parameter P and that m_P is its measure or best estimate; m_P will be thought as the best available approximation of a_P.

Let's suppose now that the higher (or lower) is the value of the parameter *P*, the higher is the related partial score; to obtain a safer pollution potential assessment because of uncertainty, then, a parameter level higher (or lower) than m_P should be used, because the resulting score will be higher than the one calculated using m_P directly. However, due to the above-mentioned interpretation of m_P, numbers increasingly different from this last are less acceptable as approximations of the parameter actual value, so only some degree of deviation is tolerable.

A choice of compromise between two requirements must be made. The model discussed hereafter, derived from SINTACS and called *f*-SINTACS ("f" stands for fuzzy) performs such choice, given some control factors. A key one is the amplitude of the range of numbers considered to some degree acceptable approximations of the actual value a_P of a parameter *P*, provided that such degree is the highest possible for m_P under all conditions; if the amplitude increases then m_P is thought to be a less reliable approximation of a_P, and the model deals with the resulting uncertainty by allowing numbers progressively distant from the former to be considered, to some extent, acceptable approximations of a_P and potentially be used to calculate the partial score. This last, because of the inference process features, will be higher than the one computed when the choice is forced to remain nearer to m_P or to coincide with it, except for special cases like those discussed later.

F-SINTACS and SINTACS produce the same pollution potential score when m_P is considered the only acceptable approximation of a_P; the effect of uncertainty is then removed, and thus it can be quantitatively assessed by comparing the scores computed under the two different conditions.

The conclusions about hydrogeological systems on which SINTACS is based are assumed to be valid; this last, particularly suited for the Mediterranean area, being well experimented in Italy has been considered a good starting point.

This work draws on previous ones regarding DRASTIC (Cameron, E., Peloso, G. F., 2000 and 2001) and, besides discussing a somewhat different assessment method, develops the analysis of the basic concepts and improves the investigation of the model behavior and the consequences of the approach employed. If further tests give positive results a limited software release may be issued.

2. SOME INTRODUCTORY CONCEPTS ABOUT FUZZY SET AND FUZZY LOGIC

A fuzzy set S is described by a function *m* that assigns to an element *x* of a chosen reference set a membership degree $m(x)$ of *x* to S, where $0 \le m(x) \le 1$. A zero membership degree, that is $m(x) = 0$, means that *x* does not belong to S while if $m(x) = 1$ then *x* belongs to S as in the ordinary set theory. Values greater than zero and less than one means, freely speaking, that *x* belongs to *S* only to a certain extent.

As an example consider the fuzzy set *high temperature* of Fig. 1 (fuzzy set names will be written in italic type); the function domain is the interval $[0,40]$ whose elements are assumed to be the temperatures, in Celsius degrees, inside a house situated in a temperate climate area.

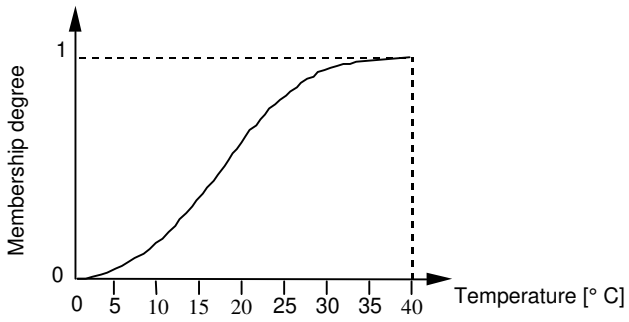


Fig. 1. The fuzzy set *high temperature*

The fuzzy set of Fig. 1 actually provides a mathematical model for the concept of high temperature and to what extent a temperature *T* is compatible with this notion is given by the membership degree $m(T)$ of *T* to the fuzzy set.

From the graph it can be observed that $g(3) = 0$, $g(40) = 1$ and $g(18) \cong 0.5$ meaning that 3, 40 and 18 degrees are, respectively, not compatible, fully compatible and half compatible with the concept of high temperature as represented by the function of Fig. 1.

The function shape is of particular importance from a semantic point of view and like other fuzzy set features depends on the specific problem and the concept to be represented. The choice, usually not univocal, may involve some subjectivity, while the final result, a function, is clearly defined; it cannot be thought as a probability density function because it has a different meaning and, in general, lacks the required properties.

Among the set-theoretical operations that can be performed with fuzzy sets intersection is of special interest for this paper: given two fuzzy sets A and B and the memberships degrees

$m_A(x)$ and $m_B(x)$ of an element *x* to each, then the membership degree of *x* to $A \cap B$ is some convenient function $m_{A\cap B}(x)$ of $m_A(x)$ and $m_B(x)$, where $0 \le m_{A\cap B}(x) \le 1$; for example $m_{A\cap B}(x) = \min[m_A(x), m_B(x)]$ for the standard Zadeh intersection operation or $m_{A\cap B}(x) = m_A(x) \cdot m_B(x)$ for the product one.

A logic based on fuzzy set theory can be developed firstly by asserting that the membership degree $m(x)$ of an element *x* to a fuzzy set S is equal to the truth degree $t(x)$ of the statement "*x* is S".

Thus, considering the preceding example, the statements P_1="[a temperature of] 3 °C is high", P_2="18 °C is high" and P_3="40 °C is high" have a truth degree $t(P_1) = m(3) = 0$, $t(P_2) = m(18) \cong 0.5$ and $t(P_3) = m(40) = 1$ meaning, again loosely speaking, that the first statement is false, the second is half true and the third is true. Phrases in square brackets will be sometimes added for clarity.

If P and Q are, respectively, the statements "*x* is A" and "*y* is B " then a truth degree $t_{P\wedge Q}$ can be assigned to the composite statement "*x* is A AND *y* is B" knowing those, $t(P)$ and $t(Q)$, of P and Q; for example $t_{P\wedge Q} = \min[t(P), t(Q)]$ or $t_{P\wedge Q} = t(P) \cdot t(Q)$ with analogy to fuzzy sets intersection. In the second case, considering the fuzzy set of fig. 1, $m(20) = 0.8$ and $m(30) = 0.9$, so the statement "20° C is high AND 30° C is high" has the truth degree $t(20) \cdot t(30) = m(20) \cdot m(30) = 0.8 \cdot 0.9 = 0.72$.

Other set-theoretical operations and logical operators are also defined but they will not be discussed here.

Fuzzy logic, naturally, includes inference methods; the simple one used by *f*-SINTACS will be exemplified later.

3. INFERENCE STEPS

The core process to assess pollution potential with *f*-SINTACS is the inference of seven weighted partial scores s^*_{Pi} $(1 \le i \le 7)$, one for each parameter considered by SINTACS, which are summed to obtain the total one $S^*_T = \sum_{i=1}^{7} s^*_{Pi}$. The partial scores are weighted because they already incorporate, as it will be clear later, the weights (the relevance) assigned to them in pollution potential assessment.

The parameters are: water table depth, net recharge, self purification effect in the zone of aeration, soil type, aquifer type with respect to lithology, hydraulic conductivity and slope of the topographic surface; the corresponding Italian words are used for the acronym SINTACS (as for DRASTIC in English).

The inference algorithm to calculate the partial scores, implemented with Wolfram Research's software Mathematica v. 4.01, will be described for net recharge and comprises the following steps:

1. Retrieval of net recharge measure or best estimate m_{NR}, though as the best approximation of the real value a_{NR}
2. Retrieval of weight w_{NR} attributed to net recharge
3. Choice of the net recharge value $\hat{v}_{NR}$ that is to the highest degree unfavourable for pollution potential and an acceptable approximation of a_{NR}, given the fuzzy set that represents this last notion.
4. Inference of the partial score s^*_{NR} through the rule "IF $\hat{v}_{NR}$ is unfavorably high AND w_{NR} is high THEN s^*_{NR} is high"

The rule (Fox E., 1994) may be read as: "insofar as the net recharge value $\hat{v}_{NR}$ is compatible with the concept of unfavorably high [for pollution potential] and as the weight w_{NR} is compatible with the concept of high, then make the partial score s^*_{NR} compatible with the concept of high".

When the process is executed using only the measure or best estimate m_{NR} of net recharge then, in the third step, $\hat{v}_{NR} = m_{NR}$.

Steps 1 to 4 are repeated for each parameter and the resulting partial pollution potential scores are added to obtain the total one.

To perform the inference it is necessary to provide the fuzzy sets that mathematically represent the concepts of acceptable approximation to net recharge actual value, unfavourably high net recharge, high weight and high partial score; also the inference method and the type of AND operator appearing in the rule, among the different options allowed with fuzzy logic, must be chosen. The first issue will be discussed in the next two paragraphs; the second will be illustrated through the exemplification.

4. THE MODEL FOR THE CONCEPT OF ACCEPTABLE APPROXIMATION TO PARAMETERS ACTUAL VALUES

Because to assess pollution potential *f*-SINTACS normally chooses parameters values different from the measures or estimates that are considered the best approximations of the actual ones, it is necessary to establish how much other options are themselves acceptable as approximations, to remain near enough to the best ones.

It is reasonable to assume that the measure or best estimate m_P of a parameter *P*, being the best approximation to the actual value a_P, fully corresponds to the concept of acceptable approximation of this last, while numbers progressively greater or smaller than m_P are less suitable; thus the fuzzy set that represents the concept of acceptable approximation of a_P should have a maximum of one in m_P and membership degrees gradually decreasing on both sides of it. Since the membership degree of m_P to the fuzzy set is one, so is the truth degree of the statement "m_P is an acceptable approximation of a_P", while for other values *x* the truth degree of the corresponding statement progressively decreases as the difference from m_P increases. It will be assumed that the membership and truth degrees are zero when $x \le m_P + e_{P1}$ or $x \ge m_P + e_{P2}$, where e_{P1} and e_{P2} are two positive values and it can be $e_{P1} \ne e_{P2}$; this means that at the ends or outside the interval $[m_P - e_{P1}, m_P + e_{P2}]$ there are no acceptable approximations of a parameter actual value a_P and therefore no possible model choices.

These notions will be discussed for net recharge.

Let us suppose that the measure or best estimate m_{NR} of net recharge for a given site is 100 mm/year and that $e_{NR1} = e_{NR2} = 25$ mm/year; the fuzzy set for the concept of acceptable approximation of net recharge actual value is a bell-shaped PI function (Cox E., 1994) which associates a membership degree of one to m_{NR} and whose domain is the interval $[m_{NR} - e_{NR1}, m_{NR} + e_{NR2}] = [100 - 25, 100 + 25] = [75, 125]$ that contains, ends excluded, all the possible model choices. The values m_{NR}, e_{NR1} and e_{NR2} univocally determine the function and thus the fuzzy set, which therefore will be denoted with $\overline{A}_{NRA}(m_{NR}, e_{NR1}, e_{NR2})$, where $\overline{A}_{NRA}$ reminds that the concept represented is of "Acceptable approximation of Net Recharge Actual value". Since m_{NR} = 100 mm/year and $e_{NR1} = e_{NR2} = 25$ mm/year the fuzzy set $\overline{A}_{NRA}(100,25,25)$ shown in fig 2 is obtained.

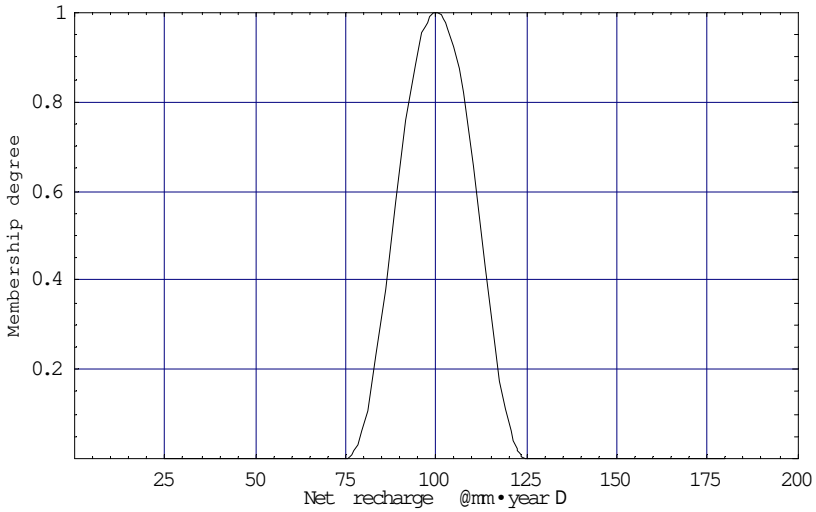


Fig. 2. The fuzzy set $\overline{A}_{NRA}(100,25,25)$

If e_{NR1} or e_{NR2} increase so does the fuzzy set domain amplitude; numbers more distant from the measure or best estimate m_{NR} are considered to some degree acceptable approximations of the net recharge actual value a_{NR}, their previous membership degree to the fuzzy set raises, possibly from 0, and they become potential model choices. This means that m_{NR} still is the best approximation of a_{NR} but is also less reliable, so more values are accepted as approximations of a_{NR} and allowed to be chosen by the model to make pollution potential assessment precautionary enough for the new uncertainty condition.

For m_{NR} = 100 mm/year and $e_{NR1} = e_{NR2} = 80$ mm/year the fuzzy set $\overline{A}_{NRA}$ becomes that of fig. 3, while an asymmetric example $\overline{A}_{NRA}(100,25,80)$ is shown in fig. 4.

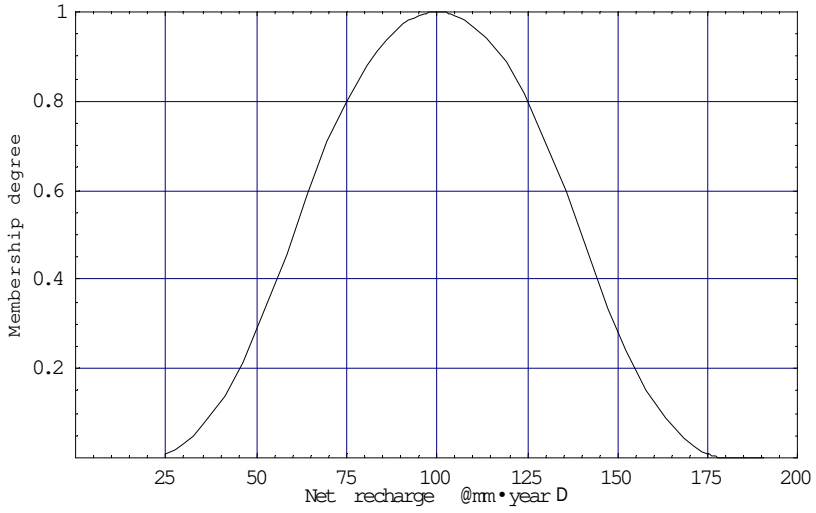


Fig. 3. The fuzzy set $\overline{A}_{NRA}(100,80,80)$

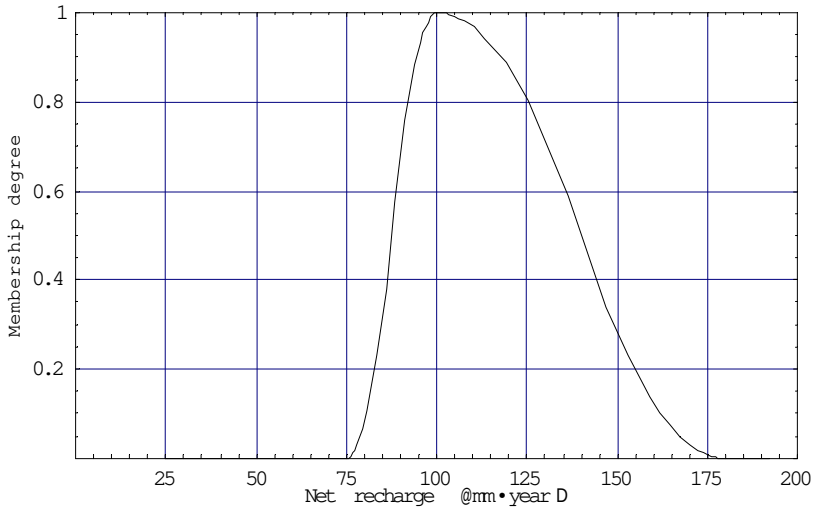


Fig. 4. The fuzzy set $\overline{A}_{NRA}(100,25,80)$

In this case the reliability of m_{NR} as an approximation to a_{NR} "from above" is considered higher then "from below".

5. THE MODELS FOR THE CONCEPTS OF UNFAVOURABLY HIGH NET RECHARGE, HIGH WEIGHT AND HIGH PARTIAL SCORE

The concept of unfavorably high net recharge is represented by the fuzzy set whose function, with domain $[25,550]$, is shown in Fig. 5.

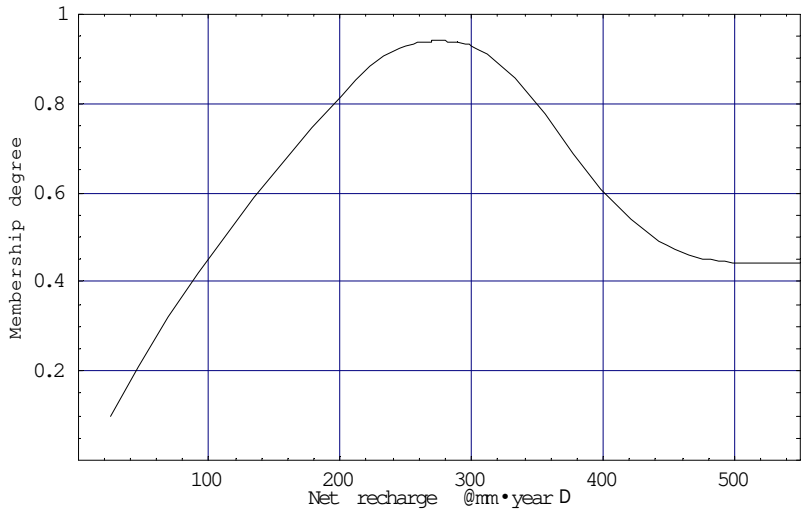


Fig. 5. The fuzzy set *unfavorably high net recharge*

The membership degree $m_1(v_{NR})$ of a given net recharge value v_{NR} to the fuzzy set expresses to what level v_{NR} can be considered unfavorably high for pollution potential, and corresponds to the truth degree $t_1[\mathrm{P}_1(v_{NR})]$ of the statement $\mathrm{P}_1(v_{NR})=$ "v_{NR} is unfavorably high"; the functional notation is used because of the variable term v_{NR}. From the graph, for example, it can be observed that the truth degree of the statement $\mathrm{P}_1(200)=$ "200 [mm/year] is unfavorably high [for pollution potential]" is $t_1[\mathrm{P}_1(200)]=m_1(200)=0.82$ while $t_1[\mathrm{P}_1(400)]=m_1(400)=0.6$.

The fuzzy set shape is that of the curve which in SINTACS directly correlates net recharge values to their respective partial scores; these latter are simply divided by 10 to obtain the membership degrees on the vertical axis. The higher the partial score, and hence the membership degree, the unfavourable the net recharge amount. Shape preservation means that the influence of net recharge on pollution potential in *f*-SINTACS and SINTACS is the same, but the mapping of Fig. 5 mathematically defines a concept for the inference algorithm.

The function values decrease after a maximum, since if an increasing net recharge favours the movement of pollutants toward groundwaters it also causes dilution and dispersion effects that lowers the concentrations; this and other phenomena reduces the adverse effect of net recharge on pollution potential beyond a value of about 280 mm/year. Also the membership degrees ranges from 0.1 to 0.95 and this limits, in *f*-SINTACS and SINTACS alike, the parameter influence on final computation.

The weights w_i (i =1, 2, …,7) attributed by SINTACS to each parameter are normally integer numbers dependent on the hydrogeological situation and the anthropic impact, but each weight can be modified by the user to better adjust the assessment to the specific case, provided that $\sum_{i=1}^{7} w_i = 26$ and $w_i \le 5$, i =1, 2, …,7.

With *f*-SINTACS how much each weight can be considered high is given by the fuzzy set named *high weight* shown in Fig. 6, a linear function whose domain is the interval $[0,\ ]$ of a discrete set; although fuzzy sets can be defined over discrete domains the choice of a continuous interval allows to treat uncertainty in weight attribution described later for parameters. This case, however, will not be discussed further.

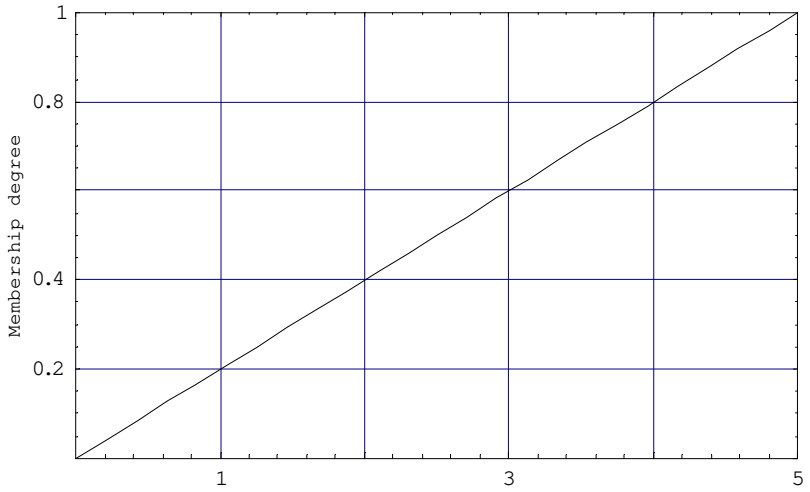


Fig. 6. The fuzzy set *high weight*

The compatibility level of a weight w_i with the concept of high weight is its membership degree $m_2(w_i)$ which corresponds to the truth degree t_2 of the statement $\mathrm{P}_2(w_i) = "w_i$ is high"; thus, for example, $t_2[\mathrm{P}_2(4)] = m_2(4) = 0.8$.

Finally *f*-SINTACS and SINTACS attributes to each parameter a partial score from 0 to 50; like for weights the mathematical model for the concept of high partial score is the linear function of Fig. 7, which can readily be interpreted considering the previous examples; thus the partial score s_i related to the *i*-th parameter has a membership degree $m_3(s_i)$ to the fuzzy set of Fig. 7 and a truth degree $t_3[\mathrm{P}_3(s_i)] = m_3(s_i)$ of the statement $\mathrm{P}_3(s_i) = "s_i$ is high".

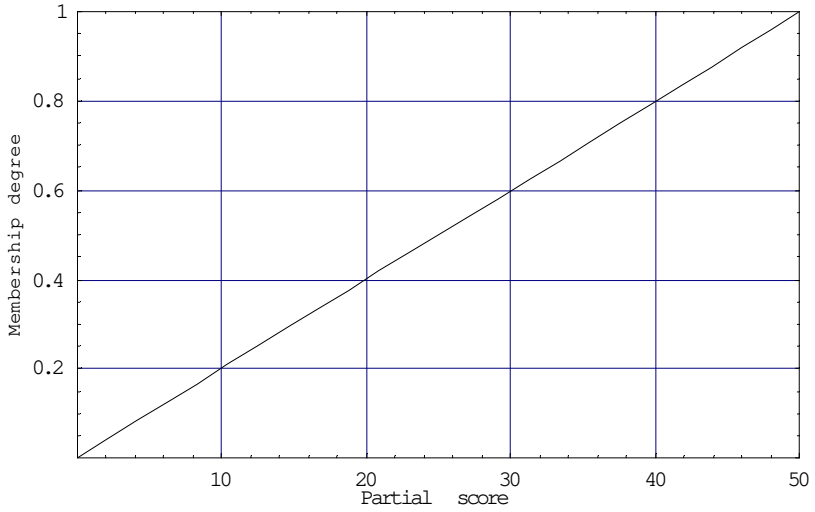


Fig. 7. The fuzzy set *high partial score*

6. PARTIAL SCORE INFERENCE USING ONLY THE MEASURE OR BEST ESTIMATE OF A PARAMETER

The partial score inference performed by *f*-SINTACS when only the measure or best estimate of a parameter is used leads to the same result as SINTACS, except possibly for small differences due essentially to numerical approximations.

The steps described in par. 3 will be exemplified for a net recharge measure or best estimate $m_{NR} = 100$ mm/year and the weight $w_{NR} = 4$ commonly assigned to the parameter.

In steps 1 and 2 the values of m_{NR} and w_{NR} are defined as inputs for the process; in step 3 $\hat{v}_{NR}$ is set equal to $m_{NR} = 100$ mm/year and the process immediately proceeds to step 4, which for clarity will be divided in sub-steps.

Sub-step 4.1. Evaluation of the truth degree t_1 of the statement $\mathrm{P}_1(100) =$ "$\hat{v}_{NR}$ is unfavorably high"; from Fig. 5, since $\hat{v}_{NR} = m_{NR} = 100$ mm/year, it is $t_1[\mathrm{P}_1(100)] = m_1(100) \cong 0.46$, where $m_1(100)$ is the membership degree of 100 mm/year to the fuzzy set *unfavorably high net recharge*.

Sub-step 4.2. Evaluation of the truth degree t_2 of the statement $\mathrm{P}_2(4) =$ "[a weight of] 4 is high"; from Fig. 6 it is $t_2[\mathrm{P}_2(4)] = m_2(4) = 0.80$, where $m_2(4)$ is the membership degree of 4 to the fuzzy set *high weight*.

Sub-step 4.3. Evaluation of the truth degree t of the composite statement $\mathrm{P}_1(100)$ AND $\mathrm{P}_2(4) =$ "[a net recharge value of] 100 [mm/year] is unfavorably high AND [a weight of] 4 is high" that forms the rule premise. The chosen AND operator ("product AND") multiplies the two truth degrees to obtain the premise one, ensures a continuous dependence of partial scores on parameters and weight values and makes those of *f*-SINTACS and SINTACS equal when uncertainty is ignored. The composite statement truth degree t is $t[\mathrm{P}_1(100) \text{ AND } \mathrm{P}_2(4)] = t_1 \cdot t_2 \cong 0.46 \cdot 0.80 \cong 0.37$.

Sub-step 4.4. Inference of the partial score s^*_{NR}. The method used, called monotonic inference, consists in searching that partial score whose membership degree to the fuzzy set *high partial score* equals the truth degree t calculated in the preceding step. This is univocally identified because the mapping of Fig. 7 is one-to-one and, since from sub-step 4.3 it is $t \cong 0.37$ a partial score $s^*_{NR} \cong 18.4$ is obtained.

The entire inference process is depicted in Fig. 8.

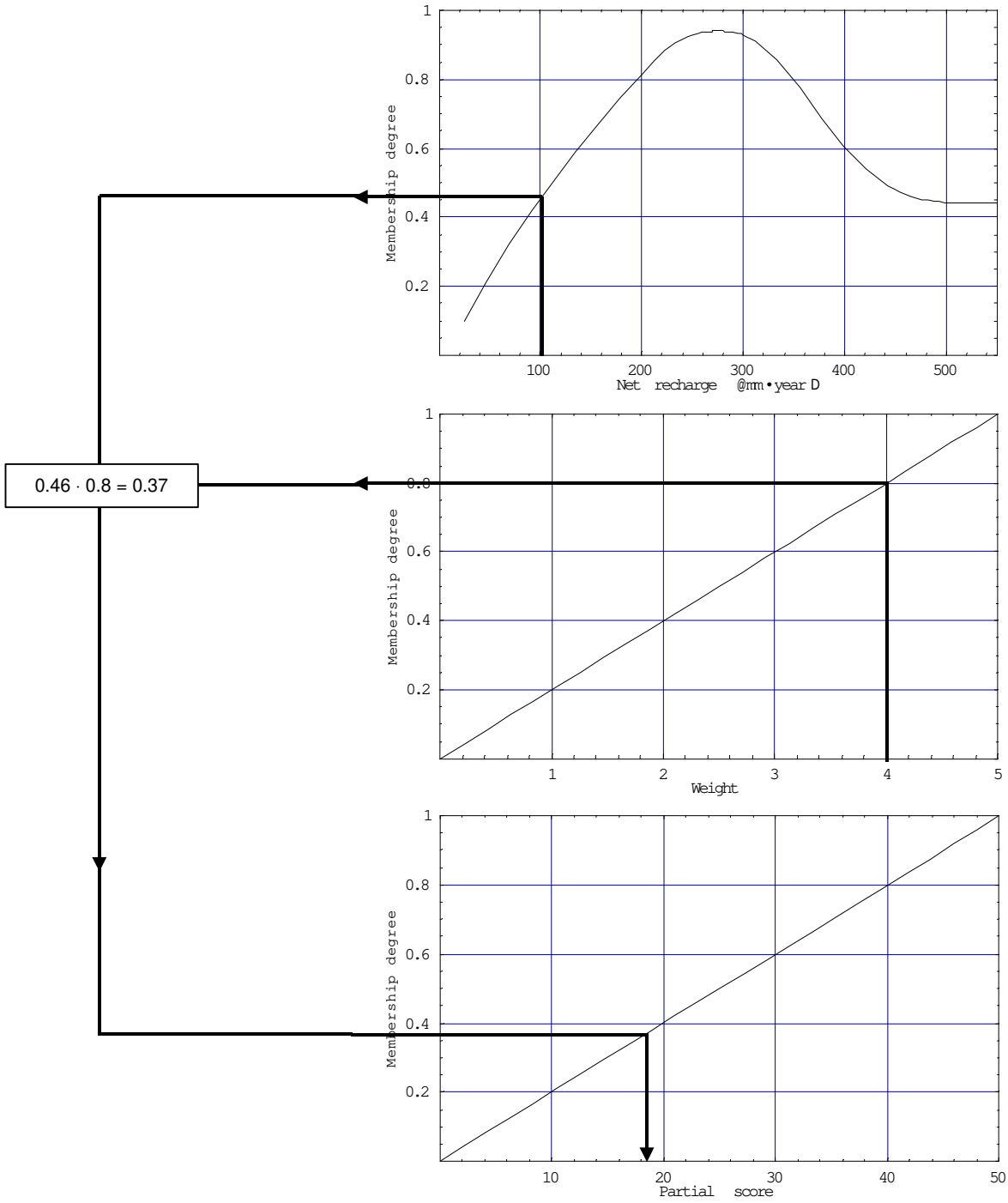


Fig. 8. Net recharge partial score inference when only the measure or best estimate of the parameter is used. Here this last is 100 mm/year and the parameter weight is 4.

7. PARTIAL SCORE INFERENCE CONSIDERING UNCERTAINTY

Because the measure or best estimate m_P of a parameter P is only the best approximation to the actual value a_P, and therefore the knowledge of this last is to some extent uncertain, *f*-SINTACS performs a precautionary pollution potential assessment by choosing a parameter value $\hat{v}_p$ that is to the highest degree both an acceptable approximation of a_P and unfavourably high for pollution potential, given the fuzzy set that represents these concepts. The model, in other words, attempts to prevent pollution potential underestimation due to uncertainty by choosing $\hat{v}_p$ so as to obtain a score higher than the one related to m_P, with a choice freedom restricted by the request that $\hat{v}_p$ must be to the highest possible extent a satisfactory approximation of a_P.

The choice is made in inference step 3 (see par. 3) using the fuzzy set representing the concept of acceptable approximation described in par. 4.

Suppose once more that the measure or best estimate of net recharge actual value a_{NR} is m_{NR} = 100 mm/year and that both the numbers outside the interval $[100-25,100+25] = [75,125]$ and its ends cannot be accepted as approximations of a_{NR}; thus the model for the concept of acceptable approximation to net recharge actual value is given by the fuzzy set $\overline{A}_{NRA}(100,25,25)$ shown in fig 2 (see again par. 4) and the fuzzy set *unfavorably high net recharge* depicted in fig. 5 (see par. 5) represents the concept implied by the name as previously explained. Since the membership degree of a net recharge value v_{NR} to the intersection between the two fuzzy sets is the minimum between the memberships degrees of v_{NR} to each of them (see par. 2) by superimposing $\overline{A}_{NRA}(100,25,25)$ and *unfavorably high net recharge* the intersection fuzzy set obtained is that highlighted in Fig. 9 by the thicker line, and the value to which the maximum membership degree corresponds is $\hat{v}_{NR} \cong 113$ mm/year.

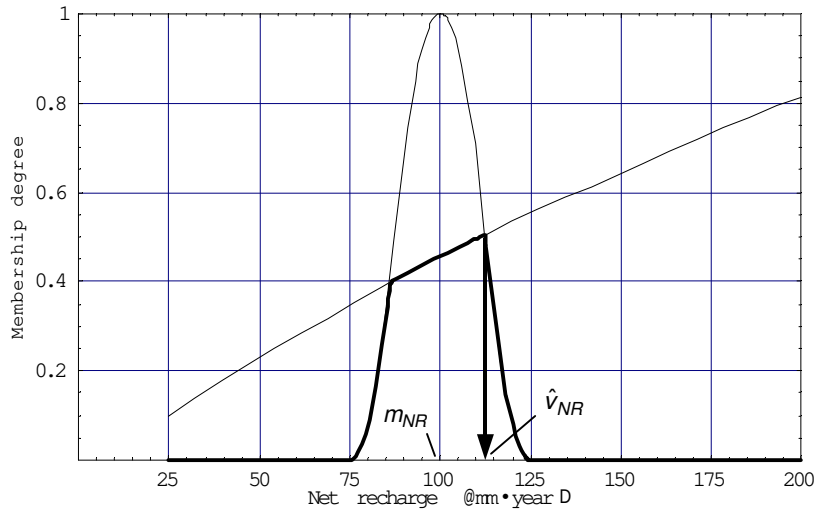


Fig. 9. Intersection between the fuzzy sets $\overline{A}_{NRA}(100,25,25)$ and *unfavorably high net recharge* (thicker line) shown for the interval $[100-25,100+25]=[75,125]$; the value $\hat{v}_{NR}$ to which the maximum membership degree corresponds is about 113.

The quantity $\hat{v}_{NR}$ has an interesting property. Values $x > \hat{v}_{NR}$ that would cause a higher pollution potential score (see fig. 5) appear to be an attractive choice for a precautionary assessment, but are less compatible than $\hat{v}_{NR}$ with the notion of acceptable approximation of net recharge actual value, because their membership degrees to the set $\overline{A}_{NRA}(100,25,25)$ are less than that of $\hat{v}_{NR}$ (see the bell-shaped curve portion that belongs to the intersection fuzzy set in fig. 9). On the other hand if $m_{NR} \le x < \hat{v}_{NR}$ the correspondence of *x* to the notion of acceptable approximation increases but the resulting partial (and hence total) score would be lower than the one obtained using $\hat{v}_{NR}$ (see fig. 9 and 10) and therefore the assessment would be less safe. Finally values below m_{NR} clearly cannot be accepted whether with SINTACS or *f*-SINTACS. Thus $\hat{v}_{NR}$, that can be regarded as an optimal compromise between acceptability and safeness, is chosen by the model to assess pollution potential. The same conclusion can be reached observing that $\hat{v}_{NR}$ maximizes the truth degree of the statement $\mathrm{P}(x)$="*x* is unfavourably high for pollution potential AND *x* is an acceptable approximation of net recharge actual value" where the AND operator is now the standard Zadeh one described in par. 2.

The inference process, represented in Fig. 10, then proceeds as explained earlier, using $\hat{v}_{NR} \cong 113$ mm/year instead of $m_{NR} = 100$ mm/year in step 4. The membership degree of $\hat{v}_{NR}$ to the fuzzy set *unfavorably high net recharge* increases from 0.46, correspondent to m_{NR}

and calculated in par. 6, to 0.51, while the partial score increases from 18.4 to 20.2 or, rounding to integer numbers, from 18 to 20 with a 11% increment.

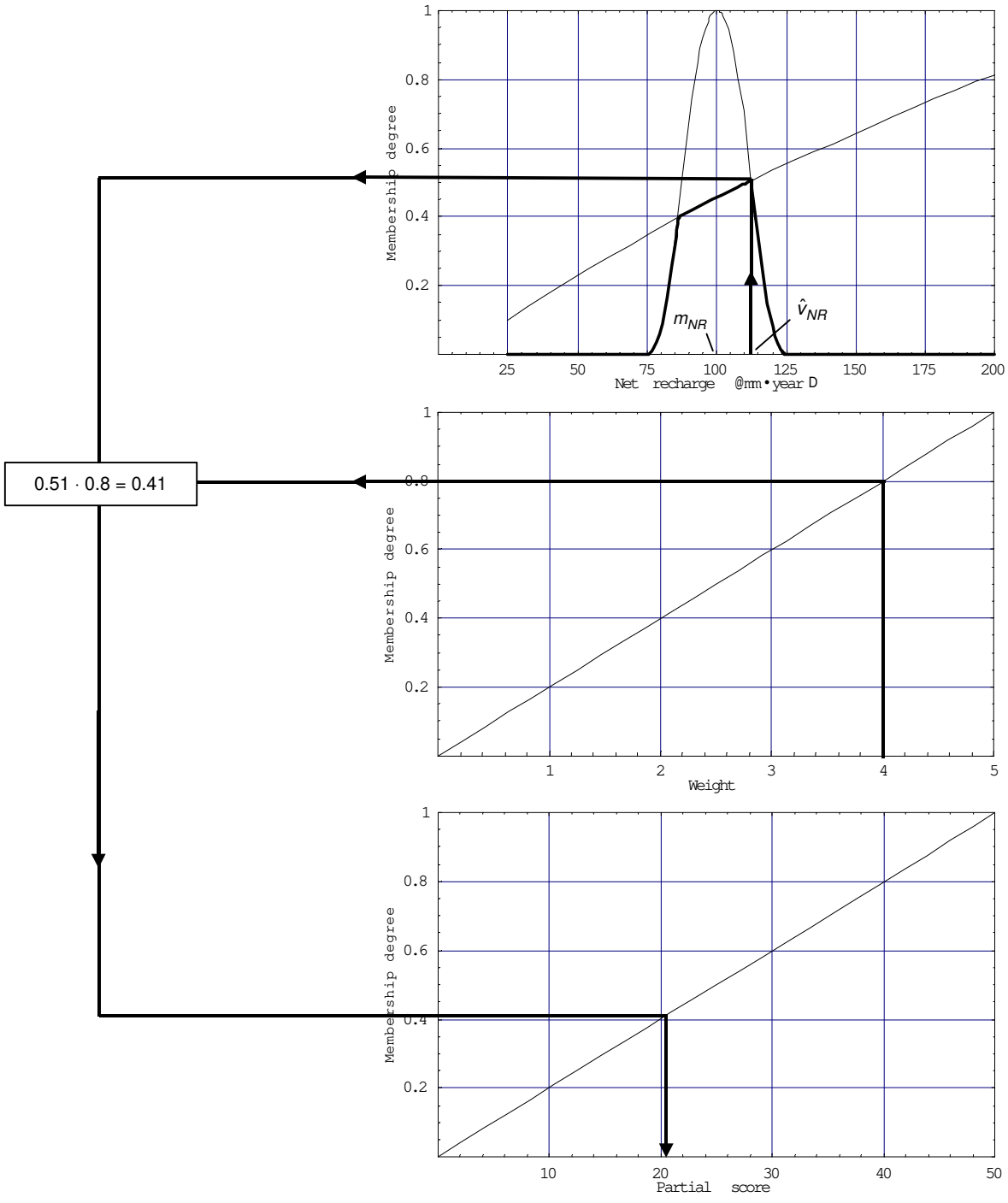


Fig. 10. Net recharge partial score inference using $\hat{v}_{NR}$. Again net recharge measure or best estimate is 100 mm/year and the parameter weight is 4; it is assumed that outside the interval $[100-25,100+25]=[75,125]$ or at its ends there are no acceptable approximation of net recharge actual value.

If the choice tolerance interval amplitude increases because the measure or best estimate m_{NR} becomes less reliable then the assessment model, in favour of safety, is allowed to select a value $\hat{v}'_{NR}$ that is more distant from m_{NR} than $\hat{v}_{NR}$ and causes and additional increment in pollution potential score; for instance if the tolerance interval is now $[100-50,100+50]=[50,150]$ the situation described in Fig. 9 becomes that of Fig. 11, $\hat{v}'_{NR} \cong 124$ mm/year (instead of $m_{NR}=100$ mm/year and $\hat{v}_{NR} \cong 113$ mm/year) and the partial score increases again from 20 to 22 with a total percentage of 22%.

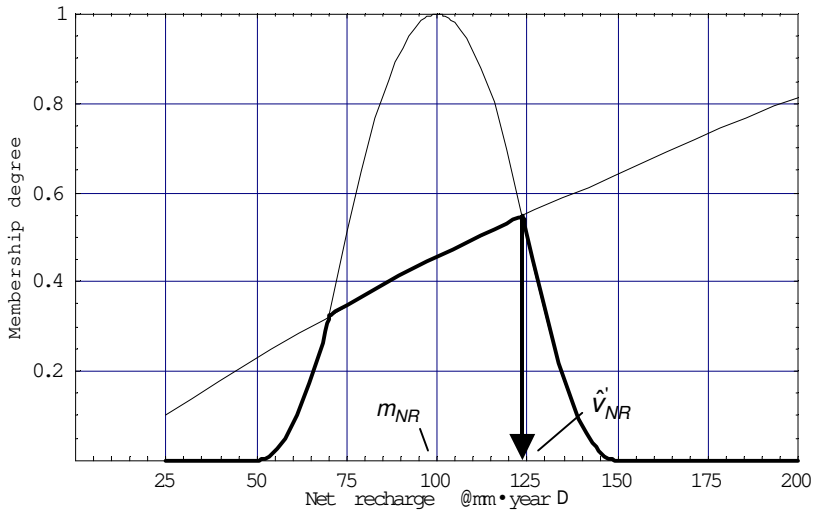


Fig. 11. Intersection between the fuzzy sets $\overline{A}_{NRA}(100,50,50)$ and *unfavorably high net recharge* (thicker line) shown for the interval $[100-50,100+50]=[50,150]$; the new value $\hat{v}'_{NR}$ to which the maximum membership degree corresponds is about 124 mm/year and the partial score increases from 20 to 22.

If m_{NR} is greater than the value to which the maximum of the fuzzy set *unfavorably high net recharge* corresponds, about 280 mm/year, then $\hat{v}_{NR} < m_{NR}$, but if the maximum belongs to the fuzzy sets intersection then $\hat{v}_{NR}$ is equal to it even if m_{NR} or the tolerance interval amplitude changes, as it is shown in Fig. 12; the adverse effect of net recharge on pollution potential and the partial score become, in fact, the highest allowed by SINTACS, and hence by *f*-SINTACS, and do not change with uncertainty nor with sufficiently small variations of m_{NR}. The model choice, moreover, cannot fall outside the permissible values range of a parameter ($[25,500]$ for net recharge) but at most corresponds to its upper or lower extreme.

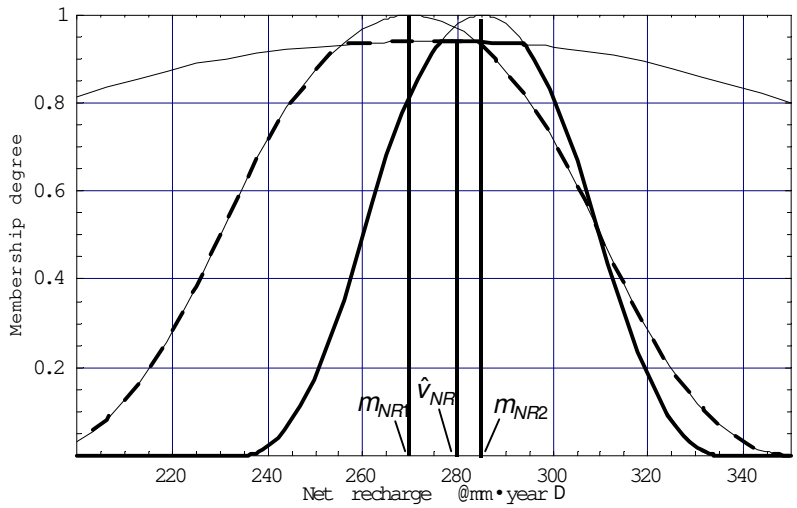


Fig. 12. If the intersection between the fuzzy sets $\overline{A}_{NRA}$ and *unfavorably high net recharge* (thicker line) contains the maximum of the latter then $\hat{v}_{NR}$ is the value, around 280 mm/year, to which the maximum is associated even if the tolerance interval amplitude changes, or m_{NR} does within certain limits

The possibility, simply by repeating the inference, to compare the results obtained ignoring or considering uncertainty as discussed, allows to quantitatively and separately determine its effect and is therefore a key model feature.

8. MODEL ANALYSIS

Decisions involving geological objects, as well as other complex systems, often imply some precaution. Foundations loads, for example, are by design normally less than soil bearing capacity also because the actual value of this latter cannot be exactly assessed due to many uncertainty factor such as soil variability, simplifications in failure phenomena description etc. Not rarely decisions involves a degree of subjectivity, which depends on experience and empirical observations.

Within *f*-SINTACS inference process subjectivity is primarily involved in choosing the functions describing fuzzy sets like $\overline{A}_{NRA}$, that are mathematical models for the concept of acceptable approximations; because the main subjective features are expressed as functions they are unambiguous, and each different proposal can be compared on a quantitative base.

Now let us assume that: 1) the risk R associated to pollution of an underground water resource is a function $R = R(P,H,V)$ of pollution potential P, hazard level H of an action and socio-economical value V of the resource, where $P,H,V \geq 0$; 2) R increases whenever one of the

variables also does, if the others remains unchanged: for example $R = P \cdot H \cdot V$; 3) P is an increasing function of pollution potential score and V is constant; 4) it is possible to define an acceptable risk level R_0 such that if $R > R_0$ the action is not undertaken or must cease and vice versa if $R \le R_0$.

Using directly the parameters measures or best estimates the same pollution potential score S_T is obtained by SINTACS and *f*-SINTACS and the corresponding risk level is $R = P(S_T) \cdot H \cdot V$; if uncertainty is considered *f*-SINTACS produces a score S_T^* where, generally, $S_T^* > S_T$ and the corresponding risk level is $R^* = P(S_T^*) \cdot H \cdot V$. Because P is an increasing function of pollution potential score if $S_T^* > S_T$ then $P(S_T^*) > P(S_T)$ and therefore $R^* > R$. To reach the acceptable risk level R_0 using both estimations it must be $R^* = R = R_0$. Since $R^* > R$ and V is constant, the hazard level when the risk is assessed considering uncertainty must be changed to a value H^* so that $R^* = P(S_T^*) \cdot H^* \cdot V = R = P(S_T) \cdot H \cdot V = R_0$ and so, dividing by V and rearranging terms, $H^* = \frac{P(S_T)}{P(S_T^*)} \cdot H$; because $P(S_T^*) > P(S_T)$ then $\frac{P(S_T)}{P(S_T^*)} < 1$ and $H^* < H$. Assessing pollution potential through *f*-SINTACS, thus, leads to a hazard reduction $\Delta H = H - H^*$ required to reach the acceptable risk level and to a strengthened pollution prevention approach because of uncertainty; the related cost $C(\Delta H)$ can be assumed to increase with ΔH. If pollution potential score is already the highest possible then $H = H^*$ and $\Delta H = 0$ regardless of uncertainty.

For the purposes of this discussion other definitions of R and the problems that can be found to specify R_0 or the functions P and, especially, H, V and C can be ignored.

Another path of thoughts can be followed. Under equal conditions if pollution potential score S_{T1}^* is obtained from more reliable parameters measures or best estimates than S_{T2}^*, normally it will be $S_{T1}^* < S_{T2}^*$. The hazard levels required to reach the acceptable risk R_0 are H_1^* and H_2^* such that $R_1^* = P(S_{T1}^*) \cdot H_1^* \cdot V = R_2^* = P(S_{T2}^*) \cdot H_2^* \cdot V = R_0$ and so, again dividing by V and rearranging terms, $H_2^* = \frac{P(S_{T1}^*)}{P(S_{T2}^*)} \cdot H_1^*$; since $S_{T1}^* < S_{T2}^*$ implies $P(S_{T1}^*) < P(S_{T2}^*)$ then $H_2^* < H_1^*$, $\Delta H_2^* = H - H_2^* > \Delta H_1^* = H - H_1^*$ and $C(\Delta H_2^*) > C(\Delta H_1^*)$, where H is the hazard level calculated without uncertainty and $H_1^*, H_2^* < H$.

Accepting a higher degree of uncertainty, then, implies an additional hazard reduction $H_1^* - H_2^*$ to reach the acceptable risk level R_0 with increased costs; this theoretically allows to compare the expenditures required to improve measurement or estimates accuracy or to reduce the action hazard level to help decide which way to proceed to achieve the risk level R_0.

Finally it is also possible to tolerate less accurate measurement or estimates, for example when it is difficult to improve data quality or in preliminary assessment, and be anyway confident that *f*-SINTACS will introduce a degree of precaution in pollution potential assessment that attempts to protect groundwaters from the potentially adverse effect of uncertainty.

The partial score s_P^* obtained by *f*-SINTACS for a given parameter *P* can be considered a function of its measure or best estimate m_P and the positive values e_{P1} and e_{P2}, which defines the fuzzy sets that are models for the concept of acceptable approximation of parameters real values, while the other fuzzy sets are fixed. It may thus be written $s_P^* = s_P^*(m_P, e_{P1}, e_{P2})$; when $e_{P1} = e_{P2} = 0$ then $s_P^*(m_P,0,0) = s_P^*(m_P) = s_P$, the score calculated by SINTACS using m_P only. The partial score $s_{NR}^* = s_{NR}^*(m_{NR}, e_{NR1}, e_{NR2})$ for a net recharge measure or best estimate $100 \le m_{NR} \le 400$ mm/year and for $10 \le e_{NR1} = e_{NR2} \le 70$ mm/year is represented by the surface of Fig. 13.

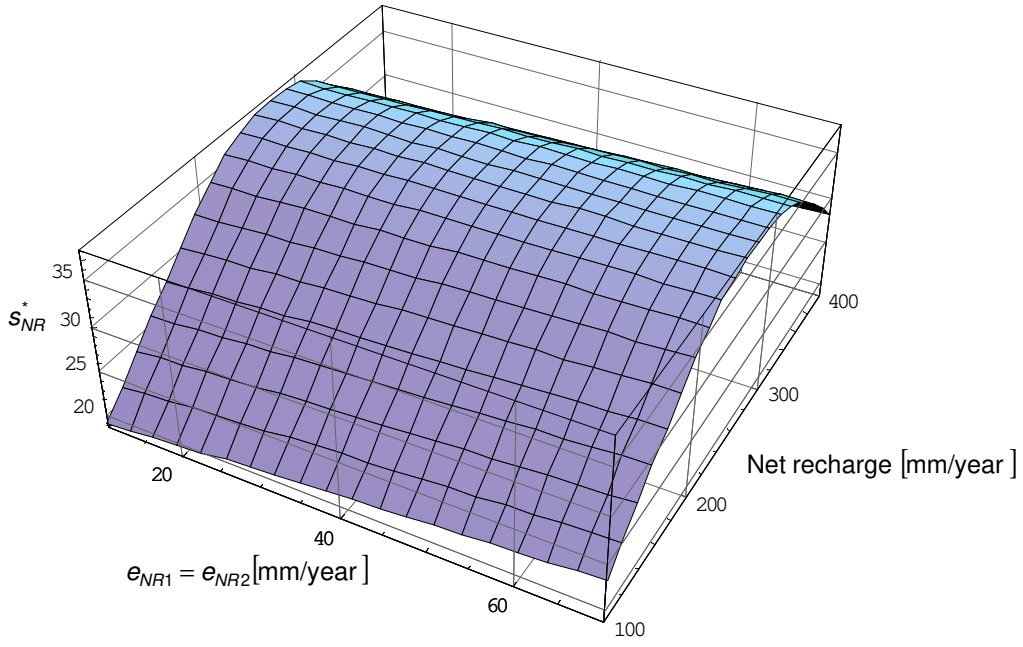


Fig. 13. The partial score $s_{NR}^* = s_{NR}^*(m_{NR}, e_{NR1}, e_{NR2})$ calculated for a net recharge measure or best estimate $100 \le m_{NR} \le 400$ mm/year and for $10 \le e_{NR1} = e_{NR2} \le 70$ mm/year.

For a given m_{NR} the partial score increases when e_{NR1} or e_{NR2} do; this means that the model introduces a degree of precaution in the assessment which intensify with uncertainty, as appropriate. The effect diminishes as m_{NR} approaches the value m_{NR0} to which the maximum of the fuzzy set *unfavourably high net recharge* corresponds, because the adverse influence of net recharge on pollution potential draw nearer to the highest possible and partial score tends to remain constant even if m_{NR}, e_{NR1} or e_{NR2} change (see par. 7); the surface top, thus, becomes progressively horizontal.

The partial score percent increment $s_i = \left(\frac{s_{NR}^{*} - s_{NR}}{s_{NR}} \right) \cdot 100$, where s_{NR} is the partial score calculated using m_{NR} only, is shown in Fig. 14, again for $100 \le m_{NR} \le 400$ mm/year and $10 \le e_{NR1} = e_{NR2} \le 70$ mm/year.

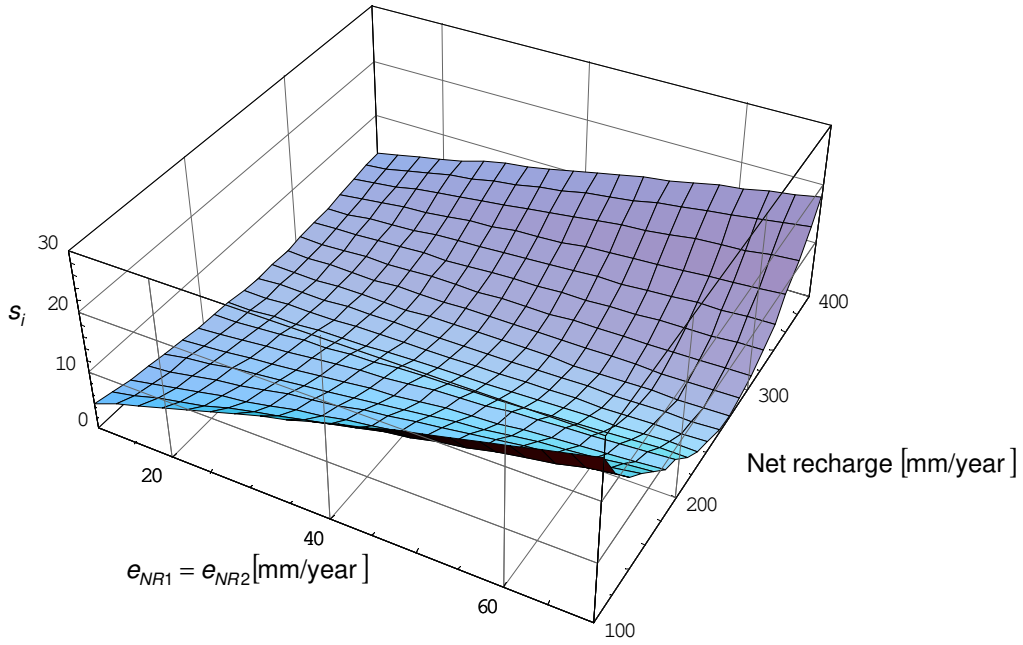


Fig. 14. The partial score percent increment $s_i = \left(\frac{s_{NR}^{*} - s_{NR}}{s_{NR}} \right) \cdot 100$ calculated for a net recharge measure or best estimate $100 \le m_{NR} \le 400$ mm/year and for $10 \le e_{NR1} = e_{NR2} \le 70$ mm/year.

The partial score percent increment taking uncertainty into account with respect to that obtained without it ranges from 0 to 26% and is already significant (over 10 %) when the precautionary choice of net recharge made by *f*-SINTACS is allowed to be within and interval of amplitude less then ± 25% of the net recharge measure or estimate m_{NR}.

The increment s_i progressively approaches 0 when m_{NR} come close to the value m_{NR0} previously considered, since s^*_{NR} tends to s_{NR} for the reasons already discussed.

The ratio absolute value $r = \left| \frac{s^*_{NR}(m_{NR}+1, e_{NR1}, e_{NR2}) - s^*_{NR}(m_{NR}, e_{NR1}, e_{NR2})}{s^*_{NR}(m_{NR}, e_{NR1}+1, e_{NR2}+1) - s^*_{NR}(m_{NR}, e_{NR1}, e_{NR2})} \right|$ quantify how sensitive is partial score to unit increments of net recharge measure or best estimate with respect to unit increments, on each side, of the length of the interval where the model choice is allowed to fall; the result, shown in fig. 15, ranges approximately between 2 and 6.

As m_{NR} approaches m_{NR0} the score s^*_{NR} becomes progressively independent from e_{NR1} and e_{NR2}, the ratio denominator tends to 0 and, after the appearance of numerical instabilities, r cannot be calculated anymore and the function surface is interrupted.

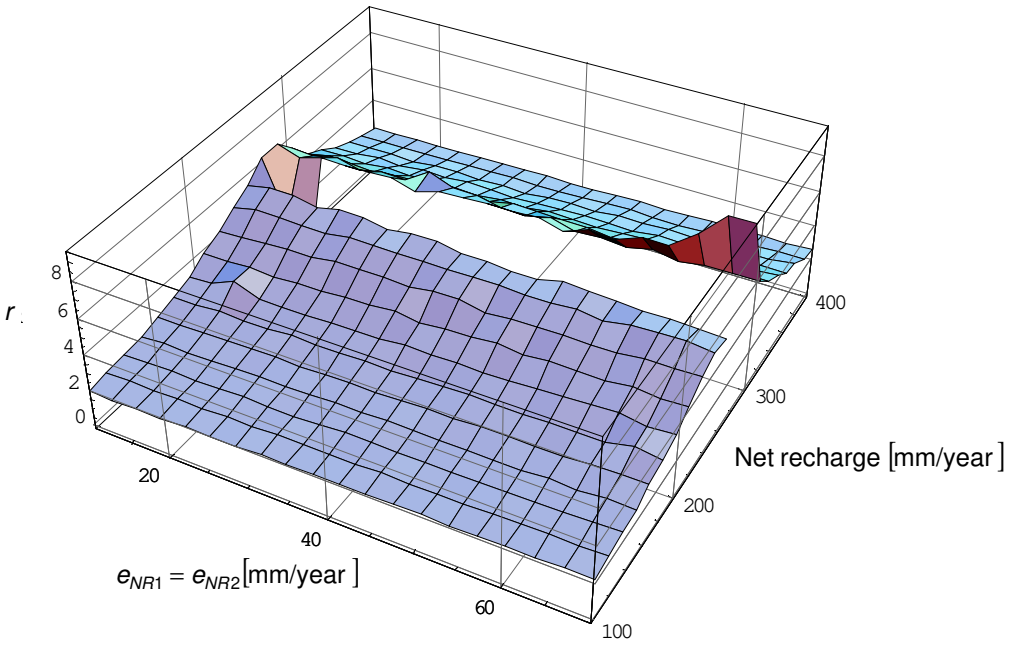


Fig. 15. The ratio absolute value $r = \left| \frac{s^*_{NR}(m_{NR}+1, e_{NR1}, e_{NR2}) - s^*_{NR}(m_{NR}, e_{NR1}, e_{NR2})}{s^*_{NR}(m_{NR}, e_{NR1}+1, e_{NR2}+1) - s^*_{NR}(m_{NR}, e_{NR1}, e_{NR2})} \right|$ calculated for a net recharge measure or best estimate $100 \le m_{NR} \le 400$ mm/year and for $10 \le e_{NR1} = e_{NR2} \le 70$ mm/year.

By comparing fig. 14, 15 and 16 it can be stated that: 1) partial score increases with uncertainty; 2) the increment is never very big even when uncertainty is considerable; 3) the increment is more sensitive to variations of parameters measures or best estimates than of uncertainty, at least when they are both small (or comparable, as it can be supposed considering the functions shapes).

These are all desirable properties. The first ensures a degree of protection induced by *f*-SINTACS that increases with uncertainty, an appropriate behavior given the model purpose. The second allows to expect that the price paid for the additional groundwater protection due to uncertainty will not be excessive. The third suggests that pollution potential score variations, under opportune conditions, depend primarily on those of parameters measures or estimates, which give the fundamental physical description of an hydrogeological system; at the same time if their reliability diminish, the importance of uncertainty in the assessment increases.

The model may suggest to protect, say, site A more than site B if parameters measure or best estimates have a lower adverse effect on pollution potential, but a higher uncertainty, in the former than in the latter; a similar situation is shown for net recharge in fig. 16.

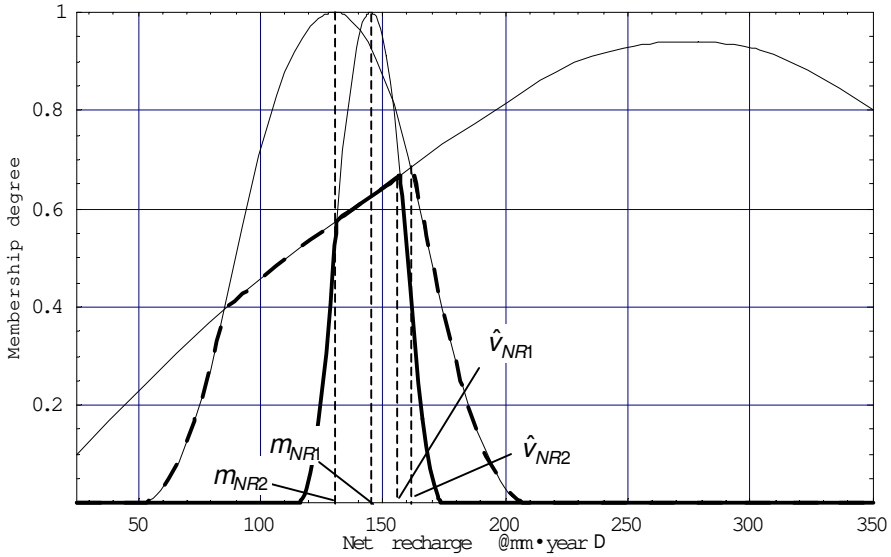


Fig. 16. Effect of differences in net recharge measures or estimation and uncertainty between two sites; it is $m_{NR1} > m_{NR2}$, but the model choices are such that $\hat{v}_{NR1} < \hat{v}_{NR2}$

This is because the model aim is to provide a shield against uncertainty itself, which is strengthened when this latter increases; the situation described, which can be easily detected by comparing pollution potential scores computed with or without uncertainty between any two sites, is expected to occur when there is a rather small difference in measures or best estimates and a marked one in uncertainty (see again fig. 16).

If parameters measures or best estimates in sites A and B lead to the same pollution potential scores and have an equal degree of uncertainty the model may suggest to protect A more than B if the effect of uncertainty can be considered more dangerous in the former than in the latter; this is when the curves portions of the fuzzy sets which express how unfavourably high or low for pollution potential a parameter is, are steeper for the values corresponding to site A than for

those of site B, since then a parameter variation can have a higher adverse effect in the former than in the latter and so does uncertainty; see fig. 17.

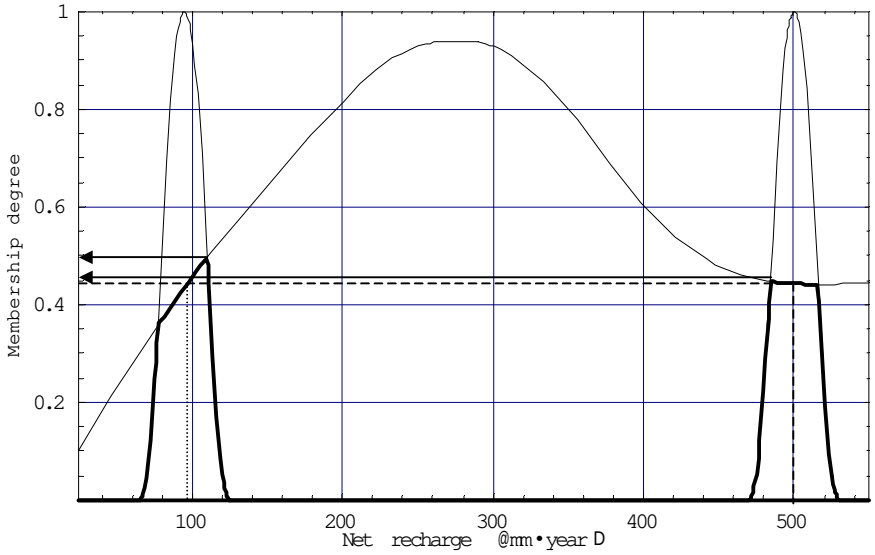


Fig. 17. Effect of parameter curve slope and uncertainty. Two net recharge values (vertical dashed lines) have the same membership degree (horizontal dashed line), and hence lead to the same partial score (see fig. 10); uncertainty is given by the bell-shapep fuzzy sets of equal amplitude. The membership degrees of the intersection maximums are different (arrows) and so are the respective pollution partial scores (see again fig. 10).

However the degrees of uncertainty should in this case be compared to parameters measures or best estimates; in fig. 17 these latter, for example, are about 95 and 500 mm/year while the bell-shaped fuzzy sets amplitude is 30 mm/year on each side in both cases. So, relatively speaking, uncertainty is much more on the first value (a 32% admissible variation) than on the second (6%). Comparing results that, in the sense described, have the same degree of uncertainty should practically eliminate this effect, which is also a consequence of slope changes as the one previously discussed.

As said before a key model feature is the possibility to repeat pollution potential score inference using only parameters measures or best estimates or considering uncertainty, obtaining two pollution potential scores S_T and S_T^* that can be compared; uncertainty effect, thus, can be quantitatively and separately evaluated for example starting from the differences $S_T^* - S_T$ at each site.

Also the model outcome should be sufficiently stable even if, to represent the concept of acceptable approximation, instead of the bell-shaped PI functions similar ones are used, such as the triangular shapes commonly employed in fuzzy set based models.

9. UNCERTAINTY DEFINITION

Because the partial score s_P^* calculated by *f*-SINTACS for a given parameter *P* is a function of its measure or best estimate m_P and the positive values e_{P1} and e_{P2} the question of how to define e_{P1} and e_{P2} arises.

There are no fixed solutions to this problem; e_{P1} and e_{P2} may, for example, be such that outside the interval $[m_P - e_{P1}, m_P + e_{P2}]$ there is a sufficiently low probability (possibly 0) to find the parameter real value, or the choice can be based on different criteria, considering also the availability and reliability of the information required to define e_{P1} and e_{P2} on a quantitative base.

10. FURTHER TESTS AND FINAL REMARKS

The model will be tested further on simulated and real hydrogeological systems, artificially introducing or estimating uncertainty about parameters values, and within decision procedures and cost/benefit schemes; the problem of how to best define e_{P1} and e_{P2} from the features of a given data set will also be analyzed in depth.

F-SINTACS may also contribute to describe, discuss and implement within a scientific framework the so-called “precaution principle” as to the debate around it.

11. REFENCES

1. Aller, L., Bennet, T., Lehr, J. H., Petty, R. J., Hackett, G. (1987). DRASTIC. A standardized system for evaluating groundwater pollution potential using hydrogeological settings. North Wales Water Authority/EPA Serv. EPA 600/2 – 85/018. Washington, 163 pp.
2. Cameron, E., Peloso, G. F. (2001). An application of fuzzy logic to the assessment of acquifers’ pollution potential. Environmental Geology, Vol. 40, num. 11/12, pp. 1305-1315. Revised and translated from the original paper published in Atti Ticinensi di Scienze della Terra, Vol. 41/2000, Ed. New Press. Como, pp. 97-108
3. Civita, M., De Maio, M. (1997). SINTACS. Un sistema parametrico per la valutazione e la cartografia della vulnerabilità degli acquiferi all’inquinamento. Metodologia e automatizzazione. Pitagora Ed. Bologna, 191 pp.
4. Cox, E. (1994). The fuzzy system handbook. Academic Press. London, 607 pp.

A NEW METHOD OF RESOURCE ESTIMATION FOR BAUXITE AND OTHER SOLID MINERAL DEPOSITS

Gy. Bárdossy [a], I. R.Szabó [b]

G. Varga [c]

a h4750bar@helka.iif.hu Hungarian Academy of Sciences, Budapest, Hungary

b Bakony Bauxite Mining Co. Tapolca, Hungary.

c Bakony Bauxite Mining Co. Tapolca, Hungary.

Abstract

First the basic concepts of resource estimation are discussed. All existing traditional methods of resource and reserve estimation are based either on deterministic or on stochastic approaches and they apply real (crisp) numbers as input data. The main disadvantage of the traditional estimation methods is that the uncertainty of the estimates cannot be determined quantitatively, instead so called „resource categories" are distinguished, e.g. measured, indicated, inferred resources.

The authors of this paper investigated new, uncertainty oriented mathematical methods, e.g. interval analysis, fuzzy set theory, probability-bounds analysis and hybrid arithmetic to calculate quantitatively the errors of the resource estimations. As a first step, they analyzed the sources of uncertainties of resource estimations, followed by a short presentation of the above listed mathematical methods. The evaluation of the errors starts at the level of the input data and the propagation of errors is followed up to the end results.

The authors performed a number of test calculations on karst bauxite deposits of Hungary, applying the fuzzy set theory. The step-be-step methodology is described and is demonstrated on figures. This methodology is relative simple and does not require high-level mathematical skill. It is suitable for the resource estimation of any other type of solid mineral deposits. The mineable and the economic reserves can be also calculated by the method recommended above. Finally the estimation results can serve as a base for risk calculations of mining investments as well.

Keywords: resource estimation, uncertainties, errors, fuzzy set theory

1. INTRODUCTION

The estimation of resources and reserves of solid mineral deposits – fossil fuels, ores and non-metallic mineral commodities – is a subject of paramount importance for mineral exploration, mining, for investors and shareholders. Hundreds of articles and several books have been published on the subject during the last five decades. General recommendations have been elaborated for the classification of resources and reserves by a commission of the United Nations (1997), including the definitions of mineral resources and reserves. Nevertheless, several open questions remained unsolved, particularly concerning the uncertainties of the estimates and the risks related to them. The problem is particularly valid to the bauxite deposits, as mining experiences often showed significant deviations from the originally calculated resources.

During the last decades new „uncertainty oriented" mathematical methods have been elaborated and successfully applied in medicine, health, biology, ecology, communication systems, economy etc. Bárdossy and J.Fodor (2000, 2001a and 2001b) discussed the general aspects of their application to geological investigations, and they carried out a number of test calculations. Bárdossy and B.Fodor (2001) discussed the possibility of applying these methods to resource and reserve estimations, in general. Bárdossy, R.Szabó and Varga (2001) carried out a number of resource estimations by these methods on different types of karst-bauxite deposits in Hungary.

In the following first the basic concepts of resource estimation, the advantages and the limitations of the traditional methods are discussed. This is followed by a short presentation of the new mathematical methods, illustrated by test calculations, carried out on selected Hungarian bauxite deposits. The estimation of *reserves* requiring the calculation of economic parameters, such as the internal rate of return (IRR) and the net present value (NPV) is beyond the scope of this article.

2. BASIC CONCEPTS OF RESOURCE ESTIMATION

Resource estimation involves the calculation of tonnage and grade of the ore (averages and spatial distribution). Different *cut-off values* are used in the different bauxite regions of the world, based on geological, mining, extraction and economic aspects of the given region. We do not discuss these aspects in this article, as they are well known and main-ly site specific.The ore resource estimation is based on *input data* obtained by the mineral exploration: mapping, trenching, pitting, drilling and geophysical measurements. Samples are taken from the deposits. The complicated problems of bauxite ore sampling have been discussed in detail by Bárdossy and Aleva (1990 pp.520-529).

For all types of resource estimation - traditional and new ones – the right geological knowledge of the deposit is a basic precondition. This knowledge is expressed in the form of a deposit *model.* This important requirement has been neglected in the past in several cases, leading to a complete misestimation of the given resources. In the case when the deposit model cannot be established unequivocally – that is two or more variants are possible – it is recommended to perform several resource estimations, one for each possible model. The results can be compared and subjective *probabilities* can be attached to them. These probabilities express the experiences of the exploration geologist and his rational opinion about the chances of each variant.

Even in the case of only one valid deposit model, the results of the resource estimation will not correspond exactly to the natural reality. The larger are the deviations from the real values, the lower is the overall *reliability* of the resource estimation. From the beginning, one of the main goals of exploration geologists was the determination and quantification of these uncertainties, designated also by the rather vague term *geological assurance.* This is the point where our new methods essentially differ from the traditional ones and offer better solution. Let us review at first the types of uncertainties and errors related to resource estimations. (For us the term uncertainty expresses the imperfection and incompleteness of our knowledge. On the other hand, *error* is the difference between a true value and an estimate of that value. In the following these two terms will be used in this sense). Two main types can be distinguished:

- *Natural variability*, an inherent feature of all geological objects and processes, is valid for the mineral deposits as well. The higher the variability of the variables included into the resource estimation, such as geometry of the deposit and grade distribution, the more uncertainties are connected with the results of the calculations. But there is a further influencing factor: in all deposits „structured" and „unstructured" features can be distinguished. The structured features, called also *trends,* can be described mathematically by trend-surface-analysis, thus they do not increase significantly the error of the resource estimation. On the other hand, local unstructured features may occur unexpectedly and their spatial position and magnitude cannot be exactly predicted. The proportion of structured and unstructured features is quite different in the mineral deposits of the world, according to our personal experiences.
- *Uncertainties related to the mineral exploration* are due to incomplete geological experience, lack of time and money and to human errors and incompetency. Their main types are as follows:
    - *Lack of representative sampling* may be the result of technical, temporal and financial restrictions. It is a significant source of uncertainty of the resource estimations.
    - *Errors of laboratory measurements („analytical errors")* comprise all chemical, mineralogical, technological and other laboratory measurements related to the resource estimation process. They consist of random and systematical components. Being studied and investigated by many authors, we do not discuss them in detail (Day, Underwood 1991).
    - *Conceptual and model uncertainties.* When evaluating a mineral deposit, existing geological concepts are applied necessarily. Unfortunately, they are not always adequate to the given deposit. Natural analogues, broadly applied in mineral exploration, are often imperfect, as they cannot take into account unknown, undetected local features. Even in the case of a valid deposit model, simplifications may increase the general error of the resource estimation.
    - *Errors due to incorrect use of mathematical methods.* They are rather frequent, as it is often neglected that resource estimation is a mathematical (statistical) procedure. And when doing so, the mathematical rules should be strictly respected. Some examples:
        - Interpolation between neighboring boreholes beyond the range of influence of the given variable may lead to erroneous results.
        - The generally calculated *arithmetic means* are valid only to normal or quasy-normal distributions. However, in mineral deposits the distribution of ore thickness and of some chemical components is often highly

skewed. In such cases *„robust" estimators* can be applied instead of the arithmetic mean. Due to this error up to 30 % we found overestimations of the bauxite resources in several former exploration reports on bauxite deposits of Hungary.

- So called „point-estimates" are reported in most cases for the deposit, instead of the mathematically required „interval-estimate" including confidence limits and levels of significance. Theoretically, the 90 and 95 % levels of significance are most suitable for the resource estimations.

- *Errors related to the choice of mathematical models.* The choice between existing mathematical models is often difficult and can be a source of additional errors. E.g. the mathematical modeling of the „experimental variograms" influences the length of the corresponding ranges of influence. The different robust estimators, mentioned above also result in different values for the measure of central tendency.

Natural variability being an inherent feature of Nature, can be studied, mathematically described, but not diminished. On the other hand, all the uncertainties and errors of mineral exploration are human factors and it depends on us how much we wish and try to diminish them.

3. ADVANTAGES AND LIMITATIONS OF THE TRADITIONAL METHODS

Two main groups of traditional resource estimations can be distinguished:

1. The scalar-geometric methods
2. The spatial („geostatistical") methods

The well known scalar-geometric methods comprise the *block methods* (triangular, regular, squares, rectangular, polygonal and irregular or geological blocks), *the methods of profiles* (vertical, horizontal, inclined) and the *isopach method.* All these methods being well known, we do not discuss them.

The principles of the *spatial methods* were elaborated by Matheron (1963) in his „theory of regionalized variables", called also *geostatistics.* This theory takes into account the spatial autocorrelation of the geological variables. By the method of *variography* the ranges of influence of the variables can be determined in two or three dimensions and predictions can be done for the spatial continuity of the deposit. An optimized estimation procedure was elaborated for the prediction of spatial points („point kriging") and for blocks („block kriging") by the solution of a set of linear equations. This new theory represented an important step ahead for resource estimations. It has been applied to bauxite deposits first in France (Maréchal and Roullier 1970), followed by Hungary (Bárdossy et al. 1985). Later the methodology has been mathematically developed and is broadly applied in all branches of mineral exploration. More than hundred articles appeared on the subject. Here we refer only to the books of Journel and Huijbregts (1978), Cressie (1991) and Goovaerts (1997). The different kriging methods use the *kriging variance* to express the uncertainty of kriging results. An improvement has been suggested by Yamamoto (1999) when applying *interpo-lation variance,* a parameter depending both on data values and data geometry. Let us stress that eveb this parameter expresses only the natural variability of the given variable and not the entire uncertainty as outlined in the first part of this article.

Further limitations of geostatistics were pointed out by Diehl (1997): thus with decreasing number of boreholes (below about 20), the calculation of variograms becomes increasingly uncertain, even impossible. This is a serious limitation for the application of the method in the case of relatively small deposits. Further uncertainties are related to the choice of the appropriate *variogram model,* as mentioned above. Furthermore, with high proportion of the *nugget-effect* the variogram model becomes uncertain. The generally applied rectangular kriging blocks depict rather roughly the real contours of the deposits. According to our own experiences, by applying different lags (distance intervals) the form and parameters of the experimental variogram may change significantly. Several important spatial features, such as the position of tectonic lines, dissecting the deposits, or the contouring of the deposit cannot be satisfactorily resolved by geostatistical methods. Finally, the application of geostatistics requires considerable expertise and a solid mathematical background, as pointed out by Diehl (1997).

The results of all the traditional methods can be presented in two ways:

- *Deterministic presentation.* In this case only the weighted averages of the resource estimation are reported, based on the well-known „best estimate" or „best guess" concept. From the mathematical point of view these are *point-estimates,* giving no information on the amount of errors of the resource estimation.
- *Probabilistic (stochastic) presentation,* when standard deviations and types of distribution are also reported. But the most important difference is, that instead of simple point-estimates *interval-estimates* are calculated involving confidence intervals at levels of confidence, chosen by the experts of the resource estimation. Uncertainties are generally expressed by the well-known *resource categories,* such as inferred, indicated and measured. However this classification is not based on uncertainty-calculations, but on the judgment of one or several experts (Diehl 1997).

Theoretically the geostatistical methods are interval estimates, as the „kriging standard deviations" refer to given levels of confidence. It should be stressed however that these standard deviations express only the amount of natural variability of the given variable and not the uncertainties related to the exploration process, as outlined above.

All traditional methods are based on the *probability theory* and consequently on its fundamental axioms, elaborated by Kolmogorov (1933). The third axiom, declaring the *principle of additivity* recognizes only mutually exclusive cases. As a consequence, all probabilistic methods have to work with well-defined, sharp boundaries and mutually exclusive geological objects. No transitions are admitted! However in geology, and particularly in bauxite deposits, sharp boundaries are rare, gradual transitions with mixed features are much more frequent, let us mention only the bauxite- clayey bauxite- bauxitic clay- clay sequence. Thus this axiom represents a serious limitation for the traditional resource estimations.

Mathematical statistics, the practical realization of the probability theory, basically requires repeated experiments (The drilling of a bore-hole is an experiment in statistical sense). However it is practically impossible to fulfill this requirement in mineral exploration. Imagine repeating 100 times a drilling grid by small shifting and rotating of the drilling locations. This would be simply nonsensical! For this reason the error of the deposit area cannot be calculated, only approximative guesses can be produced. The estimation error is also strongly influenced by the choice of the size of the estimation blocks. The larger the block, the smaller is the estimation error. Thus the categorization of the resources can be easily manipulated by changing the size of the estimation blocks.

Finally, several geological features cannot be determined exactly; they can be described only in a semi-quantitative or even qualitative way. The traditional methods of resource estimation are not suitable for the mathematical evaluation of such data; consequent-lye they have to be excluded from the estimation procedure. Much useful information is lost this way.

4. THE NEW, UNCERTAINTY ORIENTED MATHEMATICAL METHODS FOR RESOURCE ESTIMATION

All the above-discussed traditional methods apply *real numbers* (called also *crisp numbers*) as input data. However, real numbers do not express the uncertainties related to them. Together with the limitations discussed above, this is the main reason why traditional resource estimations cannot determine the total error of the estimation results. Instead *resource categories* were created to express at least approximately the amount of uncertainty of a given resource estimate, e.g. measured, indicated, inferred (Mc Kelvey 1986). Shortcomings of this concept have been discussed by many authors, e.g. Akin (1997), Diehl (1997), Wellmer (1985,1989).

During the last decades new mathematical methods have been developed, suitable to handle uncertainties from the beginning, that is from the level of the input data. Their common feature is that they apply different new types of numbers expressing the uncertainties related to them. The main methods are as follows:

- *Interval analysis* (Moore 1979) replaces the crisp numbers by uncertainty intervals. It is assumed that the true value is somewhere within the interval (***Figure 1***). Interval analysis lacks gradations and is the simplest method to express uncertainty through arithmetic calculations.
- The *fuzzy set theory* elaborated by Zadeh (1965) expresses uncertainty by fuzzy numbers. They represent estimates of uncertainty at different levels of possibility. Fuzzy numbers are by definition unimodal (convex) and they have to reach at least in one point the possibility level one, the full possibility. All arithmetic calculations can be carried out with fuzzy numbers. The main advantage of the fuzzy method is that prior geological experience can be incorporated into the construction of fuzzy numbers. This goal can be achieved by joint constructing of the fuzzy numbers by the exploration-geologist and by the mathematician. The method allows the evaluation of semi-quantitative and

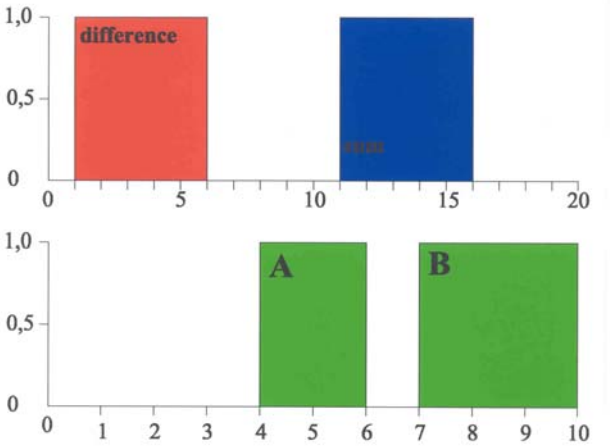
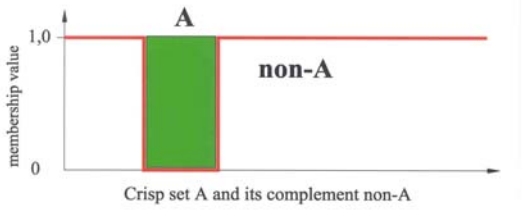
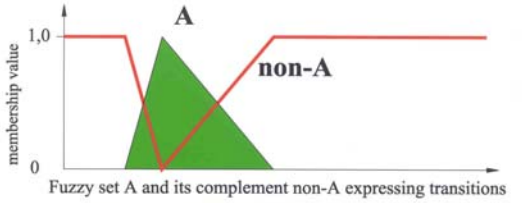


Figure 1. Two uncertainty intervals and their sum (A+B) and difference (B-A)



Crisp set A and its complement non-A



Fuzzy set A and its complement non-A expressing transitions

Figure 2. Crisp and fuzzy sets and their complements

qualitative input data as well. The transitions, mentioned above, can be also expressed by fuzzy numbers (***Figure 2***). According to our test calculations, the fuzzy numbers are highly suitable to carry out resource estimations. The development of *fuzzy geostatistic* (Bárdossy A. et al. 1990) was an essential step ahead for the evaluation of spatial uncertainty, especially for resource estimations.

- The *probability bounds theory* (Ferson et al.1999, Smith 1996, Teseem 1992) is a combination of the probability theory and the possibility theory. It expresses uncertainty by two cummulative probability distributions. The area between the two curves represents the extent of the given uncertain input data (***Figure 3***). The great advantage of this method is, that it can take into account different probability distributions, and correlations of the variables. The probability bounds get narrower with increasing information about the deposit. However the calculations are more complicated. According to our experiences, the method seems to be highly efficient for resource estimations.

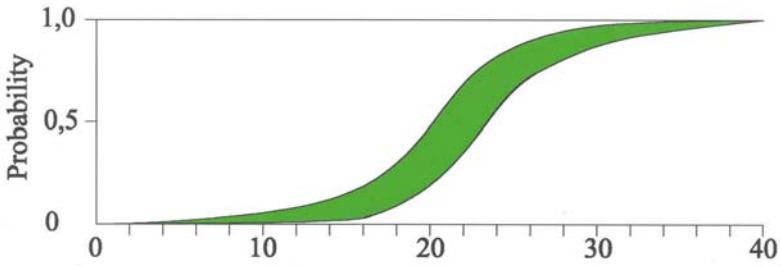


Figure 3. Uncertain input data represented by probability bounds

- The *method of hybrid arithmetic* (Cooper et al.1996, Ferson and Ginzburg 1996) combines crisp data with uncertainty intervals, fuzzy numbers and probability bounds. This is the newest among the uncertainty-oriented methods. Test calculations for resource estimations are foreseen by us in the near future.
- The methods of neural networks and fuzzy neural networks (Fullér 2001) represent useful complements to the above listed methods.

5. RESULTS OF THE TEST CALCULATIONS

For the test calculations bauxite deposits of Hungary have been chosen, as for these deposits a full and up-to-date documentation is at our disposal. It is well known that the bauxite deposits of Hungary belong to the karst-bauxite type. They are partly exploited, partly mined at present, underground and in open pits. Exploration by drilling continues at present in a number of bauxite occurrences. For the resource calculations the following *cut-off-values* are used:

Ore thickness	$\geq$ 2.0 meters
Total Al_2O_3	$\geq$ 42.0 %
Al_2O_3/ SiO_2 ratio	$\geq$ 4.0
Total sulphur as S	$\leq$ 0.6 %

In the present time the isopach and the geologic-block methods are used for resource calculations. The geostatistical method is applied in some mines only, mainly because of the lack of the required number of boreholes for a single deposit. The *input data* of these resource calculations contain the following uncertainties and errors:

- *Delineation of the contour of the deposit* (with the aim to determine the area of the deposit). The amount of the error depends mainly on the complexity of the deposit model and the spacing of the exploration grid. It is obvious, that with denser grid the

error of delineation diminishes. The calculation of the delineated area is carried out by an appropriate computer program. Its error is negligible. As repeated experiments, that is repeated drilling grids cannot be executed, the error of the deposit area cannot be exactly calculated. This is the main shortcoming of the traditional resource estimations.

- The *thickness of the ore-grade bauxite*. It is determined in the boreholes with ± 10 cm error, if the core recovery is more than 90 %. With smaller recovery the error may increase to 20-40 % for the given core interval. (In Hungarian bauxite exploration this occurs rarely). Geophysical logging can diminish this error.
- The *bulk-density is* determined by laboratory measurements of ore cores. More than 100 measurements are carried out for each deposit. The average analytical error is ±5-10 rel. %. These small-scale measurements are completed by large scale (- 1 m^3) ore samples taken from neighboring bauxite mines, assuming that they have a higher representativity.
- The average analytical errors of the *chemical analyses* are as follows: Al2O3 ±0,5 %, SiO2 ±0,3 %, CaO,MgO and total sulphur in S ±0,2 % (absolute percentages within the range of commercial bauxite composition).

Before starting the resource calculations *variograms* have been calculated by us for the bauxite-ore thickness, and for the above-mentioned chemical components, by using the well-known „VARIOWIN" computer program. All the selected bauxite deposits were checked for the drilling grid spacing: it should not surpass the calculated *ranges of influence.* This requirement was met in all cases.

From the new, uncertainty oriented methods we present here the results obtained by the fuzzy methodology. As a first step the *fuzzy numbers* of the input data have been determined. It should be stressed, that according to our experiences, the construction of the fuzzy numbers has particularities for each input variable, as it will be shown later.

An other essential difference to the traditional methods is the ranking of the boreholes. In all traditional methods only *productive* and *improductive boreholes* are distinguished. The latter ones comprise boreholes not reaching all the cut-off-values. However, according to our experience, these boreholes may also contain important information for the resource estimation, e.g. boreholes not reaching the required cut-off-thickness, but their bauxite being of high-grade composition. In other boreholes only one grade component did not reach the required value, and the thickness is also acceptable. Other boreholes contained information for the delineation of the deposit etc. We consider these cases as *transitional* and included them by the fuzzy numbers into the resource estimation. This way it was possible to increase considerably the accuracy of the final estimation results. Examples will be shown later.

The construction of the fuzzy numbers may occur in three different ways:

- For the chemical components and the bulk density the *analytical error,* established by the chemical laboratories is used, completed by the standard error of the mean and the confidence interval at 95% level of confidence.
- For semi-quantitative and qualitative variables, like the area of the deposit, the deposit model and the exploration-expert's opinion was taken into account, completed by the above mentioned direct borehole information.
- Additional estimating points have been calculated, mainly in the marginal sector to incorporate the information of the above-mentioned transitional boreholes. Here again the deposit model and the expert's opinion were taken into account

In the following the resource estimation of selected bauxite deposits will be presented: The *Szőc-Szárhegy* deposit is situated in the SW part of the Bakony Mts. It has a relatively simple

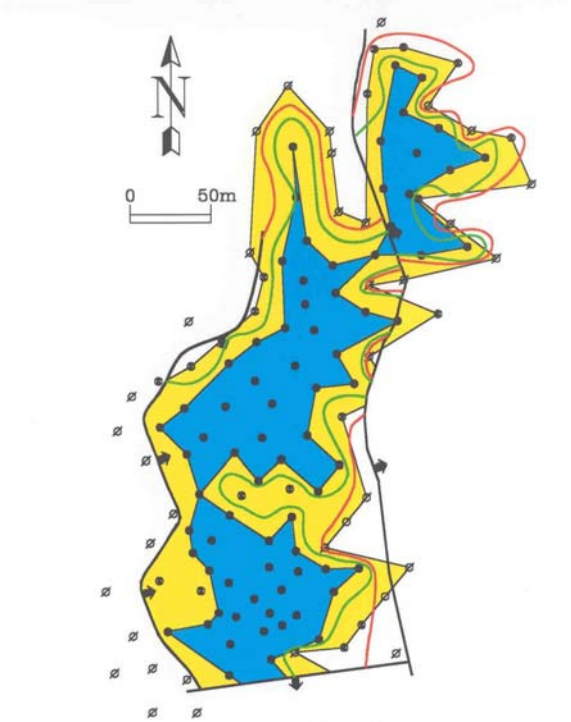


Figure 4. Resource estimation map of the Szöc–Szárhegy I. bauxite deposit

geometry: lenticular form, with the bauxite ore in its central part, surrounded laterally by clayey bauxite and bauxitic clay. The upper and lower surfaces of the ore body are relatively sharp. The ore either wedges out gradually at the rims, or finishes abruptly along fault lines. The deposit has been explored by 107 productive and 84 improductive boreholes. The resulting deposit-model has been confirmed by subsequent open-pit mining operations.

The traditional resource estimation, carried out by the geological-block method resulted in 352,7 kilotons of „geological resources". When applying the fuzzy-set theory, we have constructed fuzzy numbers for the area of the deposit, the average ore-thick-ness and the average bulk-density. Our considerations regarding the *area of the deposit* are presented on ***Figure 4.*** The smallest possible area has been constructed by connecting the marginal productive boreholes by straigt lines. The area of maximum possibility extends further, up to the closest improductive boreholes. Geological and geomorphological mapping of this near-surface deposit (overburden less than 30 meters) showed that the deposit abruptly finishes on its west side along a north-south oriented fault line, as indicated on ***Figure 4***. Here the maximum area has been diminished correspondingly to the fault line. To the south, another fault line downfaulted the bauxite that continues to the south of it. Here again the border follows the fault line. Thus the *„support"* of the fuzzy number for the deposit area extends from the minimum value of 22 261 m^2, to a maximum of 51 922 m^2. The *„core"* of the fuzzy number corresponds to the geologically most possible area, determined by a set of geological profiles and comprising 36861 m^2. The resulting fuzzy number is triangular, as presented on ***Figure 5***. A computer program, with negligible error, performed the calculation of the delineated areas.

The *average ore thickness* of the 107 productive boreholes is 4.4 m. However, the histogram showed a right-skewed distribution, confirmed also by the coefficient of skewness (1.09). In this case the simple average is biased and must be replaced by robust estimators. According to our experiences, Tukey's biweight M-estimateor has been chosen, resulting in an unbiased average of 3.9 m. The core of the corresponding fuzzy-number is an interval in this case, corresponding to the standard error of the mean, that is ± 0.2 m. Thus the core extends from 3.7 to 4.1 m. The support of the fuzzy number is longer, corresponding to the confidence interval taken at 95 % level of confidence. This interval extends from 3.6 to 4.3 m. (***Figure 5***). It should be stressed, that within both intervals there are no preferred values, as each thickness has the same possibility of occurrence.

The *average bulk-density* of the bauxite-ore has been determined by only 9 laboratory measurements. It is well known that the reliability of statistical calculations diminishes when the number of cases decreases below about 30. To eliminate this difficulty the *bootstrap method,* elaborated by Efron and Tibshirani (1993) was applied. By a computer program 1000 replicas (random samples taken by replacement) were produced and their averages calculated. The arithmetic mean of these 1000 replica-averages represents a non-biased value, in our case 2.23 g/cm^3. The analytical error of the laboratory measurements (± 5 rel. %) is included in both the core and the support of the corresponding fuzzy number. It is completed – as in the case of the thickness – by the standard error of the mean in the core, and by the confidence interval at 95 % level of confidence in the support. All these values were calculated by the above-mentioned bootstrap method to avoid bias in the results.

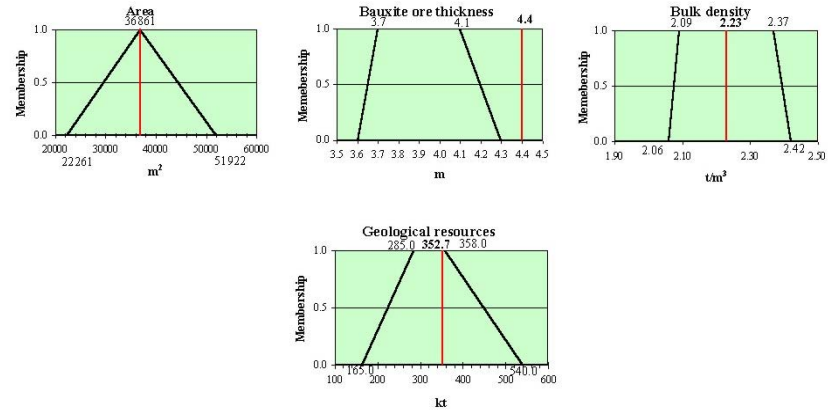


Fig.5 Fuzzy numbers of the resource calculation. Szőc-Szárhegy I. Results obtained by the "geological-block" method

The *geological resources* of the deposit are calculated by simple multiplication of the three components, when applying a traditional method. In the case of fuzzy numbers „fuzzy multiplication" has to be applied, taking into account the „error propagation". The resulting fuzzy number is shown on ***Figure 5***. Its interpretation is as follows: The minimum value of the support (165 ktons) represents the lowest possible tonnage, for the case when all components take the most unfavourable values. The maximum value of the support represents the highest possible value (540 ktons) when all components are most favorable. Both cases are theoretically possible, but their probability is close to zero. The core of the fuzzy number represents the most possible tonnage, with an interval extending from 285 to 358 ktons. The extent of this interval comprises only 73 ktons. Again, it should be stressed, that at the existing level of exploration (drilling grid), no preferred values can be chosen within this interval. But it can be declared that this interval expresses ± 11.4 % uncertainty regarding the theoretical midpoint of the interval. The tonnage calculated by the traditional method is situated within this interval (353 ktons) and is slightly biased to the right, as a consequence of the above

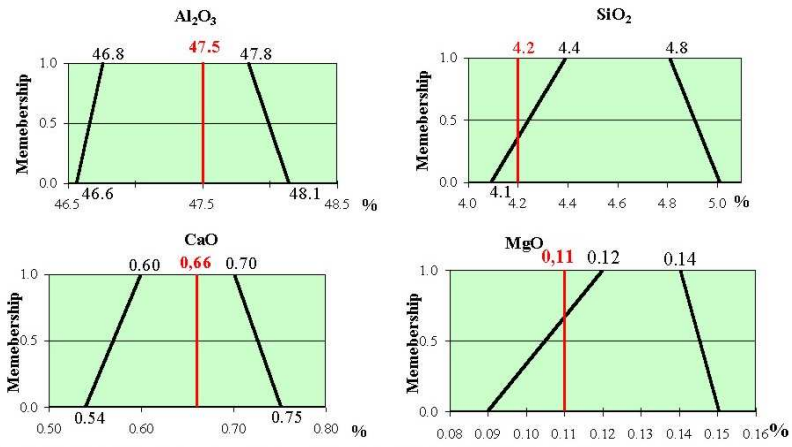
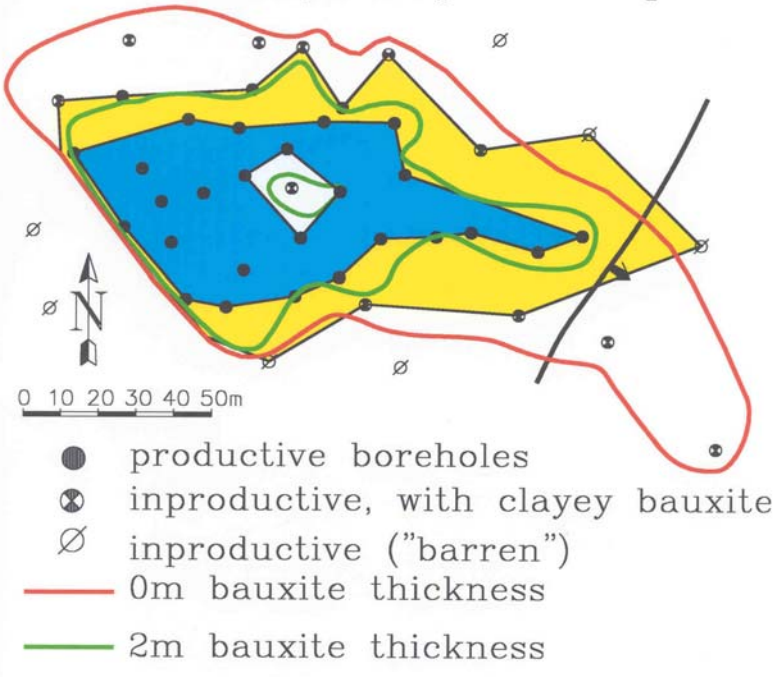


Fig.6: Fuzzy numbers of the grade calculation. Szőc-Szárhegy I. Results obtained by the "geological-block" method

mentioned biased thickness calculation (***Figure 5***).
The above-discussed fuzzy numbers clearly demonstrate the errors of the calculated resources and enable the exploration geologist and the mining engineer to take well-based decisions about the risks of a possible mining investment. The traditional results, even if they are not biased, do not give any information about these uncertainties.
The grade of the ore can be calculated in the same way. The traditional and the fuzzy numbers of the average Al_2O_3, SiO_2, CaO and MgO content are presented on ***Figure 6***. The broadness of the fuzzy numbers corresponds to the amount of the analytical error of the given component, plus the standard error of the mean in the core interval and the confidence interval at 95 % level of confidence in the support of the fuzzy number. Note how biased the traditional averages of SiO_2 and MgO are, because of the highly skewed distribution of these two chemical componets.The error of the grade calculation has been determined in the same way as that of the tonnage: it corresponds to the difference between the theoretical midpoint of the „core" and its endpoints. Thus in the case of the test deposit Szőc-Szárhegy the following results were obtained: Al2O3 ± 0.5, SiO2 ± 0.2, CaO and MgO ± 0.05, expressed as absolute percentages. These errors are fully acceptable, as they are close to the analytical error of the chemical laboratories involved.
The resources of further six bauxite deposits have been calculated by us in the same way as described above and similar experiences were obtained. Depending on the position and shape of the deposit, some additional information could be utilized for the resource estimation. E.g. observations made in the mines showed that close to vertical or very steep footwall surfaces, the bauxite becomes clayey and its brick red color changes to pink or yellow. The horizontal range of this zone is no more than 3 meters. This experience was used in sinkhole type deposits, when similar observations were made in some boreholes. At these places the the outer contour of the deposit was taken close to the given borehole, at maximum 3-4 meter distance from it. Thus the error of the determination of the deposit area could be diminished considerably (***Figure 7***).

Figure 7. Resource estimation map of the Bakonyoszlop XIII. deposit



In the Iharkút bauxite district the very complicated sinkhole type deposit of Németbánya II has been explored by a drilling grid of 10 to 20 m. The error of the area determination could be diminished by the above-mentioned method and thus the overall error of the resources is only ± 24.3 %. On the other hand, simple contours and shapes characterize the lenticular deposits of the Fenyőfő district. Despite the thinner (50 x 50 m) drilling grid our

calculations resulted in only ± 20.7 % error of the resource calculation, for the deposit No.XIII. A negative example is the Halimba II/SE deposit, characterized by particularly irregoular contours. As a consequence, a very broad area was obtained between the outer and the inner contour line. The triangular shape of the former calculations had to be changed to a trapezoidal, better expressing the increased uncertainty. Thus an overall ± 31.4 % error was obtained by the calculations, despite the close drilling grid (25 x 25 m). Further resource estimations using the outlined methodology are on the way.

6. CONCLUSIONS

- Our test calculations proved that the new method of resource calculation is relatively simple and can be carried out quickly.
- A great advantage of the method is that fuzzy numbers are robust: subjective changes do not influence significantly the results.
- By the use of fuzzy numbers the entire error of resource estimation could be determined quantitatively, this being the most important achievement of the new method.
- When using the new method, the investor can decide whether the estimation error is acceptable to him (including to the financial risk connected to it), or not. In the latter case additional exploration can be started. By the use of the suggested new method the location of the additional boreholes can be optimized and the exploration can be stopped when the required error-level is reached.
- The method outlined in this paper is suitable for any solid mineral deposits. At present further test calculations are on the way on selected lignite and building material deposits in Hungary.
- Minerable resources and commercial reserves can be calculated also by the above outlined methods.

7. REFERENCES

1. Akin,H. 1997: Zur Akzeptanz der Geostatistik als Grundlage der Vorratsklassifikation. – Schriftenreihe der GDMB.”Klassifikation von Lagerstätten” H.79.pp.9-19.
2. *Bárdossy,A.,Bogárdi,I.,Kelly,W.E. 1990: Kriging with imprecise (fuzzy) variograms. I.Theory. Mathematical Geology. 22.pp.63-79.*
3. Bárdossy,Gy. and Aleva,G.J.J. 1990: Lateritic bauxites. – Elsevier. Amsterdam-Oxford-New York-Tokyo. „Developments in Economic Geology series No.27. 624p
4. Bárdossy,Gy., Bárdossy,A., Fodor,B., Lengyel,K. 1985: Die Anwendung geostatistischer Methoden in der ungarischen Bauxitprospektion und ihr Einfluß auf den Bauxitabbau. –Erzmetall. 38. Nr.1.pp.18-22.
5. Bárdossy,Gy.and Fodor,J. 2000: Handling uncertainty in geology by new mathematical methods. – In: Proceedings Int.Symp.Hung.Researchers. „Computational Intelligence”. Budapest 2000.Nov.2. pp.93-109.
6. *Bárdossy,Gy.and Fodor,J. 2001: Experiences in the application of fuzzy sets to geology. – In:Proceedings Int.Symp.Hung.Researchers. „Computational Intelligence”. Budapest. 2001.Nov.12. pp.89-102.*

7. *Bárdossy,Gy. and Fodor,J.2001: Traditional and new ways to handle uncertainty in geology. – Natural Resources Research. Vol.10.No.3.pp.179-187.*
8. *Bárdossy,Gy.and Fodor B. 2001: New methods for the determination of uncertainties in resource estimations. (in Hungarian). – Földtani Kutatás. Budapest.No.4.pp.*
9. *Bárdossy,Gy., R.Szabó I. and Varga G. 2001: New possibilities for the evaluation of reserves as illustrated on bauxite deposits of Hungary. (in Hungarian). – Földtani Kutatás. Budapest. Vol.38.No.3.pp.35-44.*
10. *Cooper,J.A.,Ferson,S.,Ginzburg,L.R. 1996: Hybrid processing of stochastic and subjective uncertainty data. – Risk Analysis. 16.pp.785-791.*
11. *Cressie,N. 1991: Statistics for Spatial Data. – John Wiley and Sons.New York. 900p.*
12. *Day,R.A. and Underwood,P.L.1991: Quantitative analysis. – Prentice Hall,Englewood Cliff, Nwe Jersey.Sixth Edition.*
13. *Diehl,P. 1997: Quantification of the term „geological assurrance" in coal classification using geostatistical methods. – Schriftenreihe der GDMB „Klassifikation von Lagerstätten H.79.pp.187-203.*
14. *Efron,B. and Tibshirani,R.J. 1993: An introduction to the bootstrap. - Chapman &Hall New York, London.177p*
15. *Ferson,S. and Ginzburg, L.R. 1996: Different methods are needed to propagate ignorance and variability. – Reliability, Engineering and System Safety. 54.pp.133-144*
16. *Ferson,S.,Root,W.,Kuhn,R. 1999: RAMAS Risk Calc.Risk assessment with uncertain numbers. – Applied Biomathematics.Setauket, New York 184p*
17. *Goovaerts, P. 1997: Geostatistics for Natural Resources Estimation. – Oxford University Press. New York. 483*
18. *Journel,A.G. and Huijbregts,Ch.J. 1978: Mining Geostatistics. – Academic Press.New York. 600p*
19. *Maréchal, A.,Roullier,J.P.1970: Étude géostatistique des gisements de bauxite francais. – Rev.Industrie Minérale. 52. pp.492-507*
20. *Matheron,G. 1963: Principles of geostatistics. – Economic Geology.58. pp.1246-1266.*
21. *Moore,R.E. 1979: Methods and applications of interval analysis. – SIAM Studies on Applied Mathematics. Vol.2. Philadelphia.,*
22. *Smith,J.E. 1996: Generalized Chebychev inequalities: theory and applications in decision analysis. – Operations Research. 42. pp.807-825.*
23. *Tesseem,B. 1992: Interval probability propagation. – International Journal of Approximate Reasoning. 7. pp.95-120*
24. *Wellmer, F.W.1985: Rechnen für Lagerstättenkundler und Rohstoffwirtschaftler. Teil.I. –Clausthaler Tektonische Hefte.22. pp 1-187.*
25. *Wellmer,F.W. 1989: Rechnen für Lagerstättenkundler und Rohstoffwirtschaftler. Teil.II. – Clausthaler Tektonische Hefte. 26.pp 1-462*
26. *Yamamoto,K. 1999: Quantification of uncertainty in ore reserve estimation: Applications to Chapada Copper Deposit, State of Goiás, Brazil. – Natural Resources Research. Vol.8. No.1.pp153-163*
27. *Zadeh,L. 1965: Fuzzy sets. – Information and C8.pp.338-353.*

AN EXPERIMENTAL COMPARISON OF COKRIGING OF REGIONALISED COMPOSITIONAL DATA USING FOUR DIFFERENT METHODS CASE STUDY: BAUXITES IN HUNGARY

Eusebi Jarauta-Bragulat[a], Carme Hervada-Sala[b]

Angela M. Diblasi[c]

a eusebi.jarauta@upc.es Dept. de Matemàtica Aplicada III, ETSECCPB (UPC), Spain.

b carme.hervada@upc.es Dept. de Física i Enginyeria Nuclear, EUETIT (UPC), Spain.

c angelad@uncu.edu.ar Universidad Nacional de Cuyo, Argentina.

Website: ftp://ftp-urgell.upc.es/matematica/ejarauta/iamg2002/article1

Abstract

An important problem in the geosciences is the estimation or prediction of regionalized compositions. In fact, it is usual to deal with data such as percentages, concentrations, ppm,...., and use them to estimate values in other locations. Compositional data have been regarded as difficult to work with because of the so-called constant sum constraint. Following Aitchison (1986), any meaningful statement about a composition can be expressed in terms of logratios, but those transformations, and their backtransformations, are not always easy to deal with. The aim of this paper is to compare results obtained applying different methodologies developed in geostatistics, with samples of compositional data from a bauxite deposit in Halimba II (Hungary). Firstly, a classical geostatistics study is done using raw data; secondly applying two wellknown transformations in compositional data analysis: additive logratio (ALR) and centered logratio (CLR); thirdly, the Fast Fourier Transform (FFT) methodology to calculate the spatial variance-covariance matrix is used in cokriging. To be able to compare predictive values and kriging errors respective backtransforms are found. At last, results obtained with the different approaches are discussed and compared.

Keywords: Compositional regionalised data, logratios, Fast Fourier Transform, kriging.

1. INTRODUCTION

The most common goal in Geostatistics is to estimate the value of an unknown variable in a location using the information given by some samples in its surroundings. A new problem comes when compositional variables are studied, those variables are characterised by their constant sum; that is, variables summing up to one (proportions), summing up to 100 (percentages) and so on. Their main features have been studied and described by many authors (Aitchison, Barceló, Egozcue, Pawlowsky and others) and have settled some specific methodologies to work with them. Those methodologies set up some transformation of data; the best-known ones are average logratio (ALR) and centered logratio (CLR). Recently, Yao and Journel have found some way to calculate covariances matrix with Fast Fourier Transform (FFT). So, it seems sensible to apply all those methods, as well as the classical one used by most geologists to the same data to assess their applicability and results. This is the goal of this paper: the use of all those four methods, finding out their difficulties and comparing their results with some well-known data. The database is a set of compositional data from a bauxite deposit named Halimba, which is the largest one in Europe continuously mined since 1950. The data were furnished by Gy. Bárdossy, Budapest.

2. DATA SET

The studied deposit is in Hungary (Europe) and it is limited by East 117.6 - 114.0; North 13.0 - 8.8 geographic coordinates in a topographic map. The deposit covers an area of more than 8 km^2; Halimba II is the only sector in the deposit that is still under prospection. The database consists of 55 samples representing 55 boreholes, after getting off 3 incomplete samples. In these boreholes the thickness of bauxite varies from 0.8 to 36.1 m. Variables used are the following: X = Easting; Y = Northing; V_1 = Concentration of Al_2O_3; V_2 = Concentration of SiO_2; V_3 = Concentration of Fe_2O_3; V_4 = Concentration of TiO_2; V_5 = Concentration of H_2O; V_6 = Concentration of CaO; V_7 = Concentration of MgO; concentrations are in percent. The values of V_1 to V_7 represent weighted averages in each borehole taken from intervals of 0.5 to 1.0 m length. Full database and histograms of the variables can be found at our website; table 1 shows the descriptive statistics of data set.

Table 1. Descriptive statistics of data set.

	Range	Minimum	Maximum	Average	Standard deviation	simmetry	kurtosis
V1	8.3	49.9	58.2	54.569	2.234	- 0.647	- 0.558
V2	7.4	0.7	8.1	3.889	2.007	0.334	- 0.876
V3	7.4	20.4	27.8	23.698	1.898	0.523	- 0.096
V4	2.1	1.6	3.7	2.778	0.332	-0.683	2.773
V5	2.3	11.3	13.6	12.371	0.499	0.477	0.128
V6	2.7	0.1	2.8	0.536	0.545	2.232	5.543
V7	1.8	0.1	1.9	0.267	0.327	3.133	11.688

3. RAW DATA GEOSTATISTICAL ANALYSIS

This is a traditional method to estimate any regionalised variable in geostatistics. It consists of building up variograms for each variable and cross-variograms when there are more than one of them. Once experimental (cross)variograms have been built they must be modeled. The corresponding theoretical ones are used in (co)kriging system to estimate the values on a

regular grid. Variograms for the seven variables have been calculated and modeled; a table with the full description of those models can be found in our website. Once all variograms were built, cokriging has been done using KB2D program from GSLIB (1998).

4. GEOSTATISTICAL ANALYSIS CONSIDERING VARIABLES AS COMPOSITIONS: ALR TRANSFORM

Classical applications of geostatistics are related to mapping the spatial distribution of the variables under study. They give emphasis to characterize the variogram model and use the kriging (error) variance as a measure of estimation accuracy. Nowadays, some problems have been reported with compositional data. Those problems have been studied by many authors (references [1], [2], [3] and [8]). The main problem when handling compositional data is the so-called constant sum (K) constraint. Usually $K = 1$ or $K = 100$, if data are percentages. So, if V_1, ..., V_N are proportions of N elements, then $V_1 + \cdots + V_N = K$, which means that variables are not independent. To deal with compositional data and avoid this constraint, Aitchison has proposed some transforms. We have used two of them: average logratio (ALR) and centered logratio (CLR). With those transformations variables become independent and then classical kriging can be performed.

As it is said beforehand, V_1, ..., V_N must follow the constant sum constraint, but this quite never is true. Actually, we must define a new variable (called the residual) as $V_R = K - (V_1 + \cdots + V_N)$. Then, the ALR variables U_i (i = 1,2,…, N) are defined as follows:

$$U_i = alr(V_i) = \log \frac{V_i}{V_R}, i = 1,2,...,N \tag{1}$$

So now we are working with several U_i variables which do not follow the sum constraint; so they can be used as any other geostatistical data. We build and model their variograms (they can be found in our website). Once variogram was built, kriging has been done using KB2D program from GSLIB (1998). Kriging results must be backtranformed to have the estimation of V_i in the grid; in this case, the corresponding ALR-backtranform is:

$$V_i = \frac{\exp U_i}{1 + \sum_{i=1}^{N} \exp U_i} K, i = 1,2,...,N \tag{2}$$

5. GEOSTATISTICAL ANALYSIS CONSIDERING VARIABLES AS COMPOSITIONS: CLR TRANSFORM

Once V_R has been defined, we define a new variable as the geometrical mean of all of them:

$$V_{gm} = (V_1 V_2 \ldots V_N V_R)^{\frac{1}{N+1}} = \exp \frac{1}{N+1} \sum_{i=1}^{N} (\ln V_i + \ln V_R \tag{3}$$

Then, CLR transform consists in stating N+1 new variables as:

$$W_j = clr(V_i) = \ln \frac{V_j}{V_{gm}}, j = 1,2,\ldots, N+1 \quad (4)$$

These *N*+1 variables are not constrained, so they can be modeled and estimated. Once variograms have been built (they can be found in our website) cokriging has been done using KB2D program from GSLIB (1998). Then, backtransforms must be done to recover original variables. CLR-backtranform is:

$$V_j = \frac{\exp W_j}{\sum_{i=1}^{N+1} \exp W_i}, \quad j = 1,2,\ldots, N+1 \quad (5)$$

6. FAST FOURIER TRANSFORM METHOD TO CALCULATE THE COVARIANCE MATRIX

On the other hand, to avoid the modeling of variograms and crossvariograms, which may be very subjective, Yao and Journel (1998) have developed the so-called FFT method, which can be applied, in principle, to any kind of data. With FFT you do not need the independence of the variables and it builds up the covariance matrix, which can be used directly to krige. This approach works as follows:

a) Generate an experimental correlogram map on a regular grid. The grid typically has multiple nodes without estimates. The user has to specify the minimum number of data to be considered in the estimates at every node. This task is performed by program CORRMAP (see reference [6]).
b) Program INTMAP fills in the blanks typically present in the grid generated in step 1 by using a smooth local interpolation.
c) Program MULTSMTH corrects the smoothed grid to generate a third grid that is a tabulation of a positively semidefined correlogram. This condition is required to assure a unique solution for the kriging system of equations yielding a non-negative kriging variance.
d) Convert the correlogram tabulation in step 3 to covariance tabulation by multiplying the correlogram grid by the sampling variance.
e) As it was not possible to use KB2D to krige, because with this method we obtain the covariance matrix and not the variograms, we had to change it (see reference [5]).

7. RESULTS AND DISCUSSION

Table 2 shows descriptive statistics for the estimations. Variables shown are raw estimations (V_i), backtranformations of ALR estimations (BACK U_i), backtransformations of CLR estimations (BACK W_i) and estimations using FFT (FFT V_i). Table 3 shows descriptive statistics of their differences, that is the differences between raw estimations and each of the other estimations. Kriging errors can be found in website.

Figures comparing kriging results for the seven variables can be found in website; as an example you can see hereafter, in figure 1, results for variable V_1. In this figure, (a) is refered to raw data, (b) to the backtransformation of ALR-variable, (c) to the backtransformation of CLR-variable and (d) to the FFT transformation method.
Looking at the contour maps, no significant differences among the first three methods arise. However (d)-picture, the one belonging to FFT method, shows higher resolution. It seems that it is because this method is less subjective.

8. CONCLUSIONS.

Using the results of this study some conclusions can be built:

1. Kriging results in Halimba II using the four methods are quite similar.
2. As regarding to the kriging errors, comparison is not so easy because it is not true that the backtransform of ALR and CLR transformations belong to the same space as the data (this is why Martin et al. defined stress).
3. FFT method seems to be the best one, because it is less subjective, more precise and, furthermore, it is the easiest method to use. However, this method does not take into account if data are compositional or not.

9. REFERENCES

The key-rules for references format are the followings:

1. Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis simple. Proceedings of IAMG'97. IMNE, Barcelona. Part I. pp.3-35.
2. Aitchison, J. The Statistical Analysis of Compositional Data. Chapman and Hall, 1986.
3. Barceló, C. et al. Mathematical Foundations of Compositional Data Analysis. The 2001 annual conference of IAMG, Cancún (Mexico), 20 pp.
4. Deutsch, C.V. and A.G. Journel (1998). Geostatistical software library and user's guide – GSLIB. Oxford University Press, 1 CD + 369 pp.
5. Hervada-Sala, C. and E. Jarauta-Bragulat (2001). Modifications to kb2d program in GSLIB to allow use of tabulated covariances calculated with Fast Fourier Transform method. Computers & Geosciences, vol.27, num. 07, pags 887-889.
6. Ma, X and Yao,T. (2001). A program for 2D modeling (cross)correlogram tables using Fast Fourier Transform. Computers & Geosciences, vol.27, num. 07, pags 763-774.
7. Martín-Fernández, J.A. et al. (2001). Criteria to compare estimation methods of regionalized compositions. Mathematical Geology, vol 33, num. 8, pags 889-909.
8. Olea, R.A. (1999). Geostatistics for engineers and earth scientists. Kluwer Academic Publishers, 303 pp.
9. Pawlowsky, V. et al. (1995). Estimation of regionalized compositions: a comparison of three methods. Mathematical Geology, 27(1), 105–127.
10. Yao, T. and A.G. Journel (1998). Automatic modelling of (cross)covariance tables using fast Fourier transform. Mathematical Geology, 30(6), 589–615.

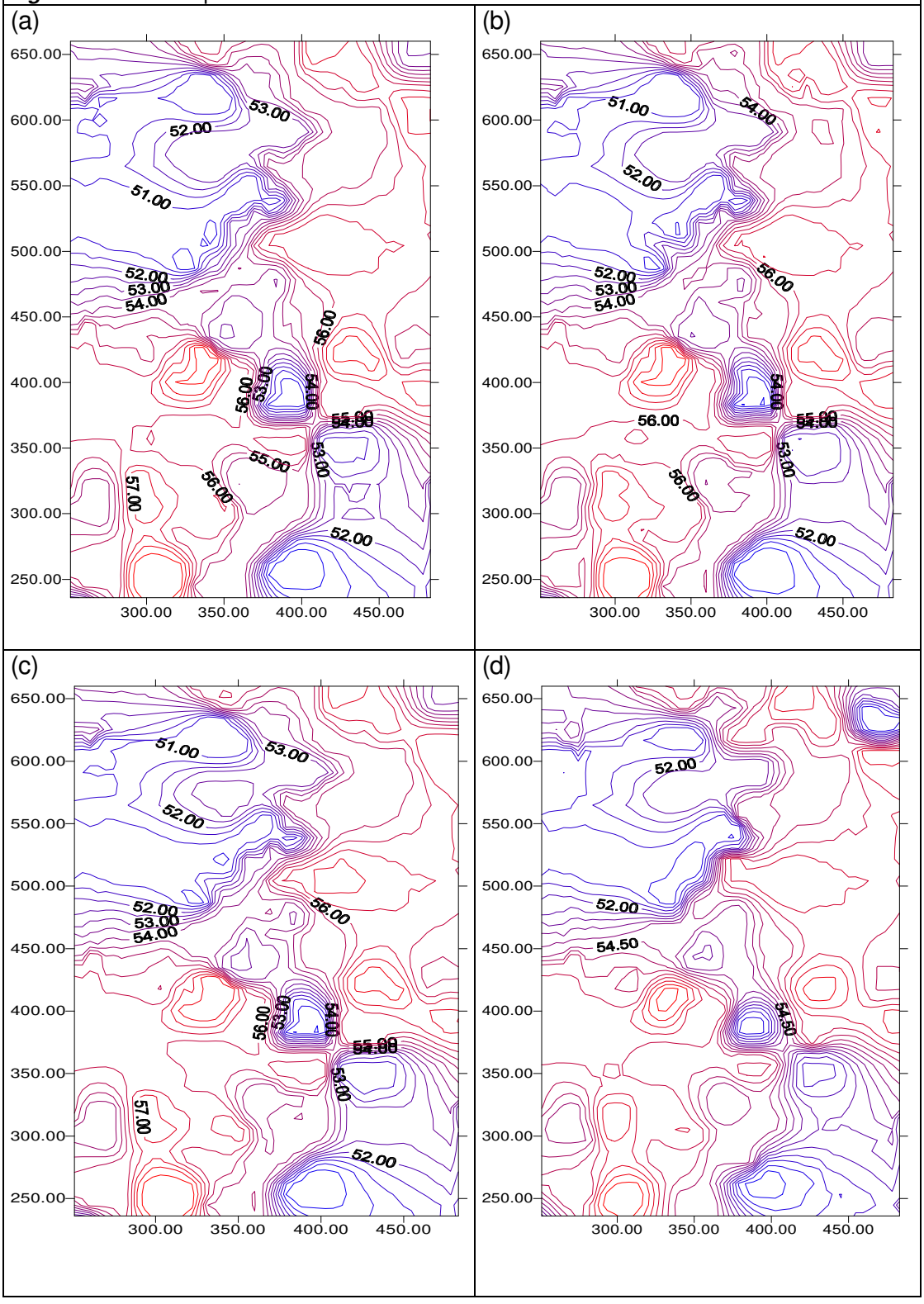
Table 2. Descriptive statistics of kriging estimations.

Variable	N	Average	Median	Std. Dev.	Minimum	Maximum
V_1	845	54.62	55.20	2.16	49.90	58.20
BACK U_1	845	54.68	55.19	2.14	49.78	58.10
BACK W_1	845	54.66	55.18	2.14	49.88	58.17
FFT V_1	578	54.62	55.30	2.24	49.90	58.20
V_2	845	3.929	3.800	2.000	0.700	8.100
BACK U_2	845	3.898	3.626	2.007	0.700	8.150
BACK W_2	845	3.899	3.623	2.006	0.700	8.150
FFT V_2	578	3.924	3.602	2.011	0.700	8.100
V_3	845	23.70	23.45	1.72	20.40	27.80
BACK U_3	845	23.72	23.41	1.72	20.32	27.75
BACK W_3	845	23.73	23.51	1.72	20.46	27.89
FFT V_3	578	23.61	23.30	1.83	20.40	27.80
V_4	845	2.777	2.800	0.348	1.600	3.700
BACK U_4	845	2.779	2.814	0.348	1.600	3.710
BACK W_4	845	2.779	2.813	0.346	1.600	3.680
FFT V_4	578	2.783	2.801	0.341	1.600	3.700
V_5	845	12.34	12.30	0.45	11.30	13.60
BACK U_5	845	12.35	12.34	0.47	11.01	13.76
BACK W_5	845	12.35	12.26	0.45	11.38	13.62
FFT V_5	578	12.37	12.30	0.50	11.30	13.60
V_6	845	0.504	0.300	0.465	0.100	2.800
BACK U_6	845	0.490	0.301	0.453	0.090	2.810
BACK W_6	845	0.490	0.301	0.453	0.100	2.810
FFT V_6	578	0.503	0.300	0.489	0.100	2.800
V_7	845	0.247	0.100	0.264	0.100	1.900
BACK U_7	845	0.234	0.105	0.247	0.090	1.900
BACK W_7	845	0.234	0.101	0.247	0.100	1.900
FFT V_7	578	0.249	0.100	0.293	0.100	1.900

Table 3. Descriptive statistics of errors.

	N	Average	Median	Std. Dev.	Minimum	Maximum
V_1 - BACK U_1	845	-0.0621	-0.0088	0.2640	-2.6162	0.6762
V_1 - BACK W_1	845	-0.0429	0.0009	0.2104	-1.7219	0.3092
V_1 - FFT V_1	560	0.0189	0.0000	0.6570	-3.1083	6.7002
V_2 - BACK U_2	845	0.0310	-0.0001	0.2165	-1.2449	2.0673
V_2 - BACK W_2	845	0.0301	-0.0011	0.1993	-1.1678	1.7569
V_2 - FFT V_2	560	0.0026	0.0000	0.4389	-3.9004	2.8890
V_3 - BACK U_3	845	-0.0148	-0.0121	0.1170	-0.9911	0.4930
V_3 - BACK W_3	845	-0.0299	-0.0262	0.1234	-1.1340	0.7786
V_3 - FFT V_3	560	0.0174	0.0000	0.6029	-3.5003	6.5000
V_4 - BACK U_4	845	-0.0018	0.0023	0.0268	-0.2413	0.2638
V_4 - BACK W_4	845	-0.0017	0.0005	0.0260	-0.1852	0.2491
V_4 - FFT V_4	560	0.0047	0.0000	0.0667	-0.6930	0.4000
V_5 - BACK U_5	845	-0.0044	0.0034	0.1598	-1.0208	1.2711
V_5 - BACK W_6	845	-0.0072	-0.0039	0.0685	-0.6018	0.2735
V_5 - FFT V_6	560	-0.0116	0.0000	0.1633	-1.0505	1.0498
V_6 - BACK U_6	845	0.0138	-0.0001	0.0633	-0.0745	0.6584
V_6 - BACK W_6	845	0.0142	0.0000	0.0681	-0.0285	0.7951
V_6 - FFT V_6	560	-0.0054	0.0000	0.1530	-1.2472	1.2500
V_7 - BACK U_7	845	0.0133	0.0001	0.0673	-0.0251	0.6202
V_7 - BACK W_7	845	0.0135	0.0001	0.0666	-0.0057	0.5916
V_7 - FFT V_7	560	-0.0040	0.0000	0.1144	-1.0331	0.9000

Figure 1. Contour maps of variable V1.



Geophysics and Geomathematics in Hungary

Laszlo Cserepes[1]

[1]Eotvos Lorand University, Department of Geophysics (laszlo.cserepes@elte.hu)

Abstract

Data processing in the above sense is now an everyday routine in most Hungarian institutions where geophysics is pursued, including university departments, research institutes, industrial laboratories and private geophysical companies. In the following description, the main emphasis will be put on those research groups which have recently contributed to the development of data processing techniques in some theoretical or methodological sense.

Keywords: geophysics, data processing.

Geomathematics, a special application of mathematical statistics to Earth science problems, aims in general at the extraction of geological objectsfrom "noisy" data sets (data including random errors). In a broad sense,this is also the basic goal of field geophysics, and this fact suggests a very close relationship between geophysics and geomathematics.

Geophysicists have long been using various mathematical methods based on probability theory, information theory and mathematical statistics in order to evaluate their field measurements. Traditionally they use the term "geophysical data processing" for this collection of data evaluation methods, but it would not be a big mistake to say "geomathematics" instead.

Notwithstanding, geophysical data processing has its own specialities, and not just because the objects of geophysical prospecting are different from, say, those of mineralogy or petrology. There are some specialities related to the methodological approach. In most geophysical data processing problems, the physical field of the geological object, which is investigated by a particular kind of geophysical measurements, is calculated from a deterministic physical theory, and these deterministically obtained theoretical values are contrasted with the stochastic data obtained in the field measurement. In this way, geophysical data processing mostly aims at an explicit physical model fitting, and the mathematical methods used in this process are selected according to this basic goal.

Data processing in the above sense is now an everyday routine in most Hungarian institutions where geophysics is pursued, including university departments, research institutes, industrial laboratories and private geophysical companies. In the following description, the main emphasis will be put on those research groups, which have recently contributed to the development of data processing techniques in some theoretical or methodological sense.

A basic problem in the theory of model fitting is the probability distribution of the "noise", i.e. the measurement and model errors. The usual "default" assumption is that these errors follow a normal (Gaussian) distribution, but it is easy to find examples where this cannot be true. This question has very important implications for the choice of the criterion function or "norm" of the model fitting. Pioneering work has been carried out in this respect by geophysicists and mathematicians working together at the Geophysical Department of Miskolc University (F. Steiner, L. Csernyák, B. Hajagos, P. Szûcs, M. Dobróka, Á. Gyulai, T. Ormos, G. Petõ, E. Turai). They have published many papers not only on the theory of establishing robust estimators, but also on actual applications of model fitting and parameter estimation for various kinds of geophysical measurements, including the joint usage of different types of measurements ("joint inversion").

Basic theory of statistical parameter estimation and examples of practical applications has long been the subject of researchers at the Geophysical Department of Eötvös University, Budapest (P. Salát, D. Drahos, K. Kis). They have focused on the classical Bayesian estimation principle, with emphasis on the use of prior information and the optimum design of measurement strategy. Important applications of the above-mentioned principles have been established for seismic and electric prospecting
by T. Fancsik, E. Prácser, G. Varga at the Eötvös Loránd Geophysical Institute of Hungary.

A very important data processing tool which, inherited from the theory of general signal processing, has become widespread in geophysics is filtering. Numerical signal filters can be designed for very different purposes of data evaluation. Data transformation by linear filters can be useful in different stages of probabilistic model fitting, and can be used simply for easing data visualization. Seismic signal processing is the largest field of filtering applications in geophysics. A. Meskó, K. Kis (Geophysical Department of Eötvös University, Budapest), F. Steiner, E. Turai (Geophysical Department of Miskolc University), L. Gömböcz, Z. Timár, P. Solt (Eötvös Loránd Geophysical Institute of Hungary), I. Késmárky, G. Göncz, I. Véges (GES Company, Budapest) are just the most important contributors to filter theory and applications.

A third large area of data evaluation which has found applications in geophysics (and in other branches of Earth sciences) is image processing. Aerial and satellite photos can be used in detecting various kinds of geophysical objects at the surface of the Earth. Besides using classical image processing tools, special geophysical applications have also been designed for this purpose. The contribution of the Space Research Group of the Geophysical Department, Eötvös University, Budapest (directed by Cs. Ferencz) and a team at the Eötvös Loránd Geophysical Institute of Hungary (directed by J. Kiss) has been valuable in this field.